

SECOND CUSTOM EDITION FOR CONCORDIA UNIVERSITY

ANALYSIS OF MARKETS

Analysis of Markets

Second Custom Edition for Concordia University

Taken from:

Microeconomics, Eighth Edition

by Robert S. Pindyck and Daniel L. Rubinfeld

The Economics of Money, Banking, and Financial Markets,
Sixth Canadian Edition

by Frederic S. Mishkin and Apostolos Serletis

Modern Labor Economics: Theory and Public Policy, Twelfth Edition
by Ronald G. Ehrenberg and Robert S. Smith

International Economics, Sixth Edition
by James Gerber

Taken from:

Microeconomics, Eighth Edition
by Robert S. Pindyck and Daniel L. Rubinfeld
Copyright © 2013, 2009, 2005, 2001 by Pearson Education, Inc.
New York, New York 10013

The Economics of Money, Banking, and Financial Markets, Sixth Canadian Edition
by Frederic S. Mishkin and Apostolos Serletis
Copyright © 2017, 2014 by Pearson Canada Inc.
Toronto, Ontario

Modern Labor Economics: Theory and Public Policy, Twelfth Edition
by Ronald G. Ehrenberg and Robert S. Smith
Copyright © 2015, 2012, 2009 by Pearson Education, Inc.
New York, New York 10013

International Economics, Sixth Edition
by James Gerber
Copyright © 2014, 2011, 2008 by Pearson Education, Inc.
New York, New York 10013

All rights reserved. No part of this book may be reproduced, in any form or by any means, without permission in writing from the publisher.

This special edition published in cooperation with Pearson Education, Inc.

All trademarks, service marks, registered trademarks, and registered service marks are the property of their respective owners and are used herein for identification purposes only.

Pearson Learning Solutions, 330 Hudson Street, New York, New York 10013
A Pearson Education Company
www.pearsoned.com

Printed in the United States of America

1 2 3 4 5 6 7 8 9 10 XXXX 19 18 17 16

000200010272042007
000200010272041524

TG

PEARSON

EBOOK ISBN 10: 1-323-42117-3
EBOOK ISBN 13: 978-1-323-42117-8



BRIEF CONTENTS

- 1** Preliminaries
- 2** The Basics of Supply and Demand
- 3** The Analysis of Competitive Markets
- 4** Overview of the Labor Market
- 5** The Demand for Labor
- 6** Supply of Labor to the Economy: The Decision to Work
- 7** Uncertainty and Consumer Behavior
- 8** An Overview of the Financial System
- 9** An Economic Analysis of Financial Structure
- 10** Understanding Interest Rates
- 11** The Behavior of Interest Rates
- 12** The Risk and Term Structure of Interest Rates
- 13** The United States in a Global Economy
- 14** Comparative Advantage and the Gains from Trade
- 15** Exchange Rates and Exchange Rate Systems

Answers to Selected & Odd Numbered Exercises at the end of this eBook

CHAPTER 1

Preliminaries

Economics is divided into two main branches: microeconomics and macroeconomics. **Microeconomics** deals with the behavior of individual economic units. These units include consumers, workers, investors, owners of land, business firms—in fact, any individual or entity that plays a role in the functioning of our economy.¹ Microeconomics explains how and why these units make economic decisions. For example, it explains how consumers make purchasing decisions and how their choices are affected by changing prices and incomes. It also explains how firms decide how many workers to hire and how workers decide where to work and how much work to do.

Another important concern of microeconomics is how economic units interact to form larger units—markets and industries. Microeconomics helps us to understand, for example, why the American automobile industry developed the way it did and how producers and consumers interact in the market for automobiles. It explains how automobile prices are determined, how much automobile companies invest in new factories, and how many cars are produced each year. By studying the behavior and interaction of individual firms and consumers, microeconomics reveals how industries and markets operate and evolve, why they differ from one another, and how they are affected by government policies and global economic conditions.

By contrast, **macroeconomics** deals with aggregate economic quantities, such as the level and growth rate of national output, interest rates, unemployment, and inflation. But the boundary between macroeconomics and microeconomics has become less and less distinct in recent years. The reason is that macroeconomics also involves the analysis of markets—for example, the aggregate markets for goods and services, labor, and corporate bonds. To understand how these aggregate markets operate, we must first understand the behavior of the firms, consumers, workers, and investors who constitute them. Thus macroeconomists have become increasingly concerned with the microeconomic foundations of aggregate economic phenomena, and much of macroeconomics is actually an extension of microeconomic analysis.

¹The prefix *micro-* is derived from the Greek word meaning “small.” However, many of the individual economic units that we will study are small only in relation to the U.S. economy as a whole. For example, the annual sales of General Motors, IBM, or Microsoft are larger than the gross national products of many countries.



CHAPTER OUTLINE

- 1.1** The Themes of Microeconomics
4
- 1.2** What Is a Market?
7
- 1.3** Real versus Nominal Prices
12
- 1.4** Why Study Microeconomics?
16

LIST OF EXAMPLES

- 1.1** The Market for Sweeteners
10
- 1.2** A Bicycle Is a Bicycle. Or Is It?
11
- 1.3** The Price of Eggs and the Price of a College Education
13
- 1.4** The Minimum Wage
15



- **microeconomics** Branch of economics that deals with the behavior of individual economic units—consumers, firms, workers, and investors—as well as the markets that these units comprise.

- **macroeconomics** Branch of economics that deals with aggregate economic variables, such as the level and growth rate of national output, interest rates, unemployment, and inflation.

1.1 The Themes of Microeconomics

The Rolling Stones once said: “You can’t always get what you want.” This is true. For most people (even Mick Jagger), that there are limits to what you can have or do is a simple fact of life learned in early childhood. For economists, however, it can be an obsession.

Much of microeconomics is about *limits*—the limited incomes that consumers can spend on goods and services, the limited budgets and technical know-how that firms can use to produce things, and the limited number of hours in a week that workers can allocate to labor or leisure. But microeconomics is also about *ways to make the most of these limits*. More precisely, it is about *the allocation of scarce resources*. For example, microeconomics explains how consumers can best allocate their limited incomes to the various goods and services available for purchase. It explains how workers can best allocate their time to labor instead of leisure, or to one job instead of another. And it explains how firms can best allocate limited financial resources to hiring additional workers versus buying new machinery, and to producing one set of products versus another.

In a planned economy such as that of Cuba, North Korea, or the former Soviet Union, these allocation decisions are made mostly by the government. Firms are told what and how much to produce, and how to produce it; workers have little flexibility in choice of jobs, hours worked, or even where they live; and consumers typically have a very limited set of goods to choose from. As a result, many of the tools and concepts of microeconomics are of limited relevance in those countries.

Trade-Offs

In modern market economies, consumers, workers, and firms have much more flexibility and choice when it comes to allocating scarce resources. Microeconomics describes the *trade-offs* that consumers, workers, and firms face, and *shows how these trade-offs are best made*.

The idea of making optimal trade-offs is an important theme in microeconomics—one that you will encounter throughout this book. Let’s look at it in more detail.

CONSUMERS Consumers have limited incomes, which can be spent on a wide variety of goods and services, or saved for the future. *Consumer theory*, the subject matter of Chapters 3, 4, and 5 of this book, describes how consumers, based on their preferences, maximize their well-being by trading off the purchase of more of some goods for the purchase of less of others. We will also see how consumers decide how much of their incomes to save, thereby trading off current consumption for future consumption.

WORKERS Workers also face constraints and make trade-offs. First, people must decide whether and when to enter the workforce. Because the kinds of jobs—and corresponding pay scales—available to a worker depend in part on educational attainment and accumulated skills, one must trade off working now (and earning an immediate income) for continued education (and the hope of earning a higher future income). Second, workers face trade-offs in their choice of employment. For example, while some people choose to work for large corporations that offer job security but limited potential for advancement, others prefer to work for small companies where there is more opportunity for



advancement but less security. Finally, workers must sometimes decide how many hours per week they wish to work, thereby trading off labor for leisure.

FIRMS Firms also face limits in terms of the kinds of products that they can produce, and the resources available to produce them. General Motors, for example, is very good at producing cars and trucks, but it does not have the ability to produce airplanes, computers, or pharmaceuticals. It is also constrained in terms of financial resources and the current production capacity of its factories. Given these constraints, GM must decide how many of each type of vehicle to produce. If it wants to produce a larger total number of cars and trucks next year or the year after, it must decide whether to hire more workers, build new factories, or do both. The *theory of the firm*, the subject matter of Chapters 6 and 7, describes how these trade-offs can best be made.

Prices and Markets

A second important theme of microeconomics is the role of *prices*. All of the trade-offs described above are based on the prices faced by consumers, workers, or firms. For example, a consumer trades off beef for chicken based partly on his or her preferences for each one, but also on their prices. Likewise, workers trade off labor for leisure based in part on the “price” that they can get for their labor—i.e., the *wage*. And firms decide whether to hire more workers or purchase more machines based in part on wage rates and machine prices.

Microeconomics also describes how prices are determined. In a centrally planned economy, prices are set by the government. In a market economy, prices are determined by the interactions of consumers, workers, and firms. These interactions occur in *markets*—collections of buyers and sellers that together determine the price of a good. In the automobile market, for example, car prices are affected by competition among Ford, General Motors, Toyota, and other manufacturers, and also by the demands of consumers. The central role of markets is the third important theme of microeconomics. We will say more about the nature and operation of markets shortly.

Theories and Models

Like any science, economics is concerned with the *explanations* of observed phenomena. Why, for example, do firms tend to hire or lay off workers when the prices of their raw materials change? How many workers are likely to be hired or laid off by a firm or an industry if the price of raw materials increases by, say, 10 percent?

In economics, as in other sciences, explanation and prediction are based on *theories*. Theories are developed to explain observed phenomena in terms of a set of basic rules and assumptions. The *theory of the firm*, for example, begins with a simple assumption—firms try to maximize their profits. The theory uses this assumption to explain how firms choose the amounts of labor, capital, and raw materials that they use for production and the amount of output they produce. It also explains how these choices depend on the *prices* of inputs, such as labor, capital, and raw materials, and the prices that firms can receive for their outputs.

Economic theories are also the basis for making predictions. Thus the theory of the firm tells us whether a firm’s output level will increase or decrease in response to an increase in wage rates or a decrease in the price of raw materials. With the application of statistical and econometric techniques, theories can be used to construct models from which quantitative predictions can be made.



A *model* is a mathematical representation, based on economic theory, of a firm, a market, or some other entity. For example, we might develop a model of a particular firm and use it to predict *by how much* the firm's output level will change as a result of, say, a 10-percent drop in the price of raw materials.

Statistics and econometrics also let us measure the *accuracy* of our predictions. For example, suppose we predict that a 10-percent drop in the price of raw materials will lead to a 5-percent increase in output. Are we sure that the increase in output will be exactly 5 percent, or might it be somewhere between 3 and 7 percent? Quantifying the accuracy of a prediction can be as important as the prediction itself.

No theory, whether in economics, physics, or any other science, is perfectly correct. The usefulness and validity of a theory depend on whether it succeeds in explaining and predicting the set of phenomena that it is intended to explain and predict. Theories, therefore, are continually tested against observation. As a result of this testing, they are often modified or refined and occasionally even discarded. The process of testing and refining theories is central to the development of economics as a science.

When evaluating a theory, it is important to keep in mind that it is invariably imperfect. This is the case in every branch of science. In physics, for example, Boyle's law relates the volume, temperature, and pressure of a gas.² The law is based on the assumption that individual molecules of a gas behave as though they were tiny, elastic billiard balls. Physicists today know that gas molecules do not, in fact, always behave like billiard balls, which is why Boyle's law breaks down under extremes of pressure and temperature. Under most conditions, however, it does an excellent job of predicting how the temperature of a gas will change when the pressure and volume change, and it is therefore an essential tool for engineers and scientists.

The situation is much the same in economics. For example, because firms do not maximize their profits all the time, the theory of the firm has had only limited success in explaining certain aspects of firms' behavior, such as the timing of capital investment decisions. Nonetheless, the theory does explain a broad range of phenomena regarding the behavior, growth, and evolution of firms and industries, and has thus become an important tool for managers and policymakers.

Positive versus Normative Analysis

Microeconomics is concerned with both *positive* and *normative* questions. Positive questions deal with explanation and prediction, normative questions with what *ought* to be. Suppose the U.S. government imposes a quota on the import of foreign cars. What will happen to the price, production, and sales of cars? What impact will this policy change have on American consumers? On workers in the automobile industry? These questions belong to the realm of **positive analysis**: statements that describe relationships of cause and effect.

Positive analysis is central to microeconomics. As we explained above, theories are developed to explain phenomena, tested against observations, and used to construct models from which predictions are made. The use of economic theory for prediction is important both for the managers of firms and for public policy. Suppose the federal government is considering raising the tax on gasoline. The change would affect the price of gasoline, consumers' purchasing

• **positive analysis** Analysis describing relationships of cause and effect.

²Robert Boyle (1627–1691) was a British chemist and physicist who discovered experimentally that pressure (P), volume (V), and temperature (T) were related in the following way: $PV = RT$, where R is a constant. Later, physicists derived this relationship as a consequence of the kinetic theory of gases, which describes the movement of gas molecules in statistical terms.



choices for small or large cars, the amount of driving that people do, and so on. To plan sensibly, oil companies, automobile companies, producers of automobile parts, and firms in the tourist industry would all need to estimate the impact of the change. Government policymakers would also need quantitative estimates of the effects. They would want to determine the costs imposed on consumers (perhaps broken down by income categories); the effects on profits and employment in the oil, automobile, and tourist industries; and the amount of tax revenue likely to be collected each year.

Sometimes we want to go beyond explanation and prediction to ask such questions as “What is best?” This involves **normative analysis**, which is also important for both managers of firms and those making public policy. Again, consider a new tax on gasoline. Automobile companies would want to determine the best (profit-maximizing) mix of large and small cars to produce once the tax is in place. Specifically, how much money should be invested to make cars more fuel-efficient? For policymakers, the primary issue is likely to be whether the tax is in the public interest. The same policy objectives (say, an increase in tax revenues and a decrease in dependence on imported oil) might be met more cheaply with a different kind of tax, such as a tariff on imported oil.

Normative analysis is not only concerned with alternative policy options; it also involves the design of particular policy choices. For example, suppose it has been decided that a gasoline tax is desirable. Balancing costs and benefits, we then ask what is the optimal size of the tax.

Normative analysis is often supplemented by value judgments. For example, a comparison between a gasoline tax and an oil import tariff might conclude that the gasoline tax will be easier to administer but will have a greater impact on lower-income consumers. At that point, society must make a value judgment, weighing equity against economic efficiency. When value judgments are involved, microeconomics cannot tell us what the best policy is. However, it can clarify the trade-offs and thereby help to illuminate the issues and sharpen the debate.

• **normative analysis** Analysis examining questions of what ought to be.

1.2 What Is a Market?

Business people, journalists, politicians, and ordinary consumers talk about markets all the time—for example, oil markets, housing markets, bond markets, labor markets, and markets for all kinds of goods and services. But often what they mean by the word “market” is vague or misleading. In economics, markets are a central focus of analysis, so economists try to be as clear as possible about what they mean when they refer to a market.

It is easiest to understand what a market is and how it works by dividing individual economic units into two broad groups according to function—*buyers* and *sellers*. Buyers include consumers, who purchase goods and services, and firms, which buy labor, capital, and raw materials that they use to produce goods and services. Sellers include firms, which sell their goods and services; workers, who sell their labor services; and resource owners, who rent land or sell mineral resources to firms. Clearly, most people and most firms act as both buyers and sellers, but we will find it helpful to think of them as simply buyers when they are buying something and sellers when they are selling something.

Together, buyers and sellers interact to form *markets*. A **market** is the collection of buyers and sellers that, through their actual or potential interactions, determine the price of a product or set of products. In the market for personal computers, for example, the buyers are business firms, households, and students; the sellers are

• **market** Collection of buyers and sellers that, through their actual or potential interactions, determine the price of a product or set of products.



• **market definition**

Determination of the buyers, sellers, and range of products that should be included in a particular market.

• **arbitrage** Practice of buying at a low price at one location and selling at a higher price in another.

• **perfectly competitive market** Market with many buyers and sellers, so that no single buyer or seller has a significant impact on price.

• **market price** Price prevailing in a competitive market.

Hewlett-Packard, Lenovo, Dell, Apple, and a number of other firms. Note that a market includes more than an *industry*. An *industry* is a collection of firms that sell the same or closely related products. In effect, an industry is the supply side of the market.

Economists are often concerned with **market definition**—with determining which buyers and sellers should be included in a particular market. When defining a market, *potential* interactions of buyers and sellers can be just as important as *actual* ones. An example of this is the market for gold. A New Yorker who wants to buy gold is unlikely to travel to Zurich to do so. Most buyers of gold in New York will interact only with sellers in New York. But because the cost of transporting gold is small relative to its value, buyers of gold in New York *could* purchase their gold in Zurich if the prices there were significantly lower.

Significant differences in the price of a commodity create a potential for **arbitrage**: buying at a low price in one location and selling at a higher price somewhere else. The possibility of arbitrage prevents the prices of gold in New York and Zurich from differing significantly and creates a world market for gold.

Markets are at the center of economic activity, and many of the most interesting issues in economics concern the functioning of markets. For example, why do only a few firms compete with one another in some markets, while in others a great many firms compete? Are consumers necessarily better off if there are many firms? If so, should the government intervene in markets with only a few firms? Why have prices in some markets risen or fallen rapidly, while in other markets prices have hardly changed at all? And which markets offer the best opportunities for an entrepreneur thinking of going into business?

Competitive versus Noncompetitive Markets

In this book, we study the behavior of both competitive and noncompetitive markets. A **perfectly competitive market** has many buyers and sellers, so that no single buyer or seller has any impact on price. Most agricultural markets are close to being perfectly competitive. For example, thousands of farmers produce wheat, which thousands of buyers purchase to produce flour and other products. As a result, no single farmer and no single buyer can significantly affect the price of wheat.

Many other markets are competitive enough to be treated as if they were perfectly competitive. The world market for copper, for example, contains a few dozen major producers. That number is enough for the impact on price to be small if any one producer goes out of business. The same is true for many other natural resource markets, such as those for coal, iron, tin, or lumber.

Other markets containing a small number of producers may still be treated as competitive for purposes of analysis. For example, the U.S. airline industry contains several dozen firms, but most routes are served by only a few firms. Nonetheless, because competition among those firms is often fierce, for some purposes airline markets can be treated as competitive. Finally, some markets contain many producers but are *noncompetitive*; that is, individual firms can jointly affect the price. The world oil market is one example. Since the early 1970s, that market has been dominated by the OPEC cartel. (A *cartel* is a group of producers that acts collectively.)

Market Price

Markets make possible transactions between buyers and sellers. Quantities of a good are sold at specific prices. In a perfectly competitive market, a single price—the **market price**—will usually prevail. The price of wheat in Kansas



City and the price of gold in New York are two examples. These prices are usually easy to measure. For example, you can find the price of corn, wheat, or gold each day in the business section of a newspaper.

In markets that are not perfectly competitive, different firms might charge different prices for the same product. This might happen because one firm is trying to win customers from its competitors, or because customers have brand loyalties that allow some firms to charge higher prices than others. For example, two brands of laundry detergent might be sold in the same supermarket at different prices. Or two supermarkets in the same town might sell the same brand of laundry detergent at different prices. In cases such as this, when we refer to the market price, we will mean the price averaged across brands or supermarkets.

The market prices of most goods will fluctuate over time, and for many goods the fluctuations can be rapid. This is particularly true for goods sold in competitive markets. The stock market, for example, is highly competitive because there are typically many buyers and sellers for any one stock. As anyone who has invested in the stock market knows, the price of any particular stock fluctuates from minute to minute and can rise or fall substantially during a single day. Likewise, the prices of commodities such as wheat, soybeans, coffee, oil, gold, silver, and lumber can rise or fall dramatically in a day or a week.

Market Definition—The Extent of a Market

As we saw, *market definition* identifies which buyers and sellers should be included in a given market. However, to determine which buyers and sellers to include, we must first determine the **extent of a market**—its *boundaries*, both *geographically* and in terms of the *range of products* to be included in it.

When we refer to the market for gasoline, for example, we must be clear about its geographic boundaries. Are we referring to downtown Los Angeles, southern California, or the entire United States? We must also be clear about the range of products to which we are referring. Should regular-octane and high-octane premium gasoline be included in the same market? Gasoline and diesel fuel?

For some goods, it makes sense to talk about a market only in terms of very restrictive geographic boundaries. Housing is a good example. Most people who work in downtown Chicago will look for housing within commuting distance. They will not look at homes 200 or 300 miles away, even though those homes might be much cheaper. And homes (together with the land they are sitting on) 200 miles away cannot be easily moved closer to Chicago. Thus the housing market in Chicago is separate and distinct from, say, that in Cleveland, Houston, Atlanta, or Philadelphia. Likewise, retail gasoline markets, though less limited geographically, are still regional because of the expense of shipping gasoline over long distances. Thus the market for gasoline in southern California is distinct from that in northern Illinois. On the other hand, as we mentioned earlier, gold is bought and sold in a world market; the possibility of arbitrage prevents the price from differing significantly from one location to another.

We must also think carefully about the range of products to include in a market. For example, there is a market for single-lens reflex (SLR) digital cameras, and many brands compete in that market. But what about compact “point-and-shoot” digital cameras? Should they be considered part of the same market? Probably not, because they are typically used for different purposes and so do not compete with SLR cameras. Gasoline is another example. Regular- and premium-octane gasolines might be considered part of the same market because

- **extent of a market**

Boundaries of a market, both geographical and in terms of range of products produced and sold within it.



most consumers can use either. Diesel fuel, however, is not part of this market because cars that use regular gasoline cannot use diesel fuel, and vice versa.³

Market definition is important for two reasons:

- A company must understand who its actual and potential competitors are for the various products that it sells or might sell in the future. It must also know the product boundaries and geographical boundaries of its market in order to set price, determine advertising budgets, and make capital investment decisions.
- Market definition can be important for public policy decisions. Should the government allow a merger or acquisition involving companies that produce similar products, or should it challenge it? The answer depends on the impact of that merger or acquisition on future competition and prices; often this can be evaluated only by defining a market.

EXAMPLE 1.1 THE MARKET FOR SWEETENERS

In 1990, the Archer-Daniels-Midland Company (ADM) acquired the Clinton Corn Processing Company (CCP).⁴ ADM was a large company that produced many agricultural products, one of which was high-fructose corn syrup (HFCS). CCP was another major U.S. corn syrup producer. The U.S. Department of Justice (DOJ) challenged the acquisition on the grounds that it would lead to a dominant producer of corn syrup with the power to push prices above competitive levels. Indeed, ADM and CCP together accounted for over 70 percent of U.S. corn syrup production.

ADM fought the DOJ decision, and the case went to court. The basic issue was whether corn syrup represented a distinct market. If it did, the combined market share of ADM and CCP would have been about 40 percent, and the DOJ's concern might have been warranted. ADM, however, argued that the correct market definition was much broader—a market for sweeteners which included sugar as well as corn syrup. Because the ADM–CCP combined share of a sweetener market would have been quite small, there would be no concern about the company's power to raise prices.

ADM argued that sugar and corn syrup should be considered part of the same market because they

are used interchangeably to sweeten a vast array of food products, such as soft drinks, spaghetti sauce, and pancake syrup. ADM also showed that as the level of prices for corn syrup and sugar fluctuated, industrial food producers would change the proportions of each sweetener that they used in their products. In October 1990, a federal judge agreed with ADM's argument that sugar and corn syrup were both part of a broad market for sweeteners. The acquisition was allowed to go through.

Sugar and corn syrup continue to be used almost interchangeably to satisfy Americans' strong taste for sweetened foods. The use of all sweeteners rose steadily through the 1990s, reaching 150 pounds per person in 1999. But starting in 2000, sweetener use began to decline as health concerns led people to find substitute snacks with less added sugar. By 2010, American per-capita consumption of sweeteners had dropped to 130 pounds per person. In addition, for the first time since 1985, people consumed more sugar (66 pounds per person) than corn syrup (64.5 pounds per person). Part of the shift from corn syrup to sugar was due to a growing belief that sugar is somehow more “natural”—and therefore healthier—than corn syrup.

³How can we determine the extent of a market? Since the market is where the price of a good is established, one approach focuses on market prices. We ask whether product prices in different geographic regions (or for different product types) are approximately the same, or whether they tend to move together. If either is the case, we place them in the same market. For a more detailed discussion, see George J. Stigler and Robert A. Sherwin, “The Extent of the Market,” *Journal of Law and Economics* 27 (October 1985): 555–85.

⁴This example is based on F. M. Scherer, “Archer-Daniels-Midland Corn Processing,” Case C16-92-1126, John F. Kennedy School of Government, Harvard University, 1992.



EXAMPLE 1.2 A BICYCLE IS A BICYCLE. OR IS IT?

Where did you buy your last bicycle? You might have bought a used bike from a friend or from a posting on Craigslist. But if it was new, you probably bought it from either of two types of stores.

If you were looking for something inexpensive, just a functional bicycle to get you from A to B, you would have done well by going to a mass merchandiser such as Target, Wal-Mart, or Sears. There you could easily find a decent bike costing around \$100 to \$200. On the other hand, if you are a serious cyclist (or at least like to think of yourself as one), you would probably go to a bicycle dealer—a store that specializes in bicycles and bicycle equipment. There it would be difficult to find a bike costing less than \$400, and you could easily spend far more. But of course you would have been happy to spend more, because you are serious cyclist.

What does a \$1000 Trek bike give you that a \$120 Huffy bike doesn't? Both might have 21-speed gear shifts (3 in front and 7 in back), but the shifting mechanisms on the Trek will be higher quality and probably shift more smoothly and evenly. Both bikes will have front and rear hand brakes, but the brakes on the Trek will likely be stronger and more durable. And the Trek is likely to have a lighter



frame than the Huffy, which could be important if you are a competitive cyclist.

So there are actually two different markets for bicycles, markets that can be identified by the type of store in which the bicycle is sold. This is illustrated in Table 1.1. "Mass market" bicycles, the ones that are sold in Target and Wal-Mart, are made by companies such as Huffy, Schwinn, and Mantis, are priced as low as \$90 and rarely cost more than \$250. These companies are focused on producing functional bicycles as cheaply as possible, and typically do

their manufacturing in China. "Dealer" bicycles, the ones sold in your local bicycle store, include such brands as Trek, Cannondale, Giant, Gary Fisher, and Ridley, and are priced from \$400 and up—way up. For these companies the emphasis is on performance, as measured by weight and the quality of the brakes, gears, tires, and other hardware.

Companies like Huffy and Schwinn would never try to produce a \$1000 bicycle, because that is simply not their forte (or competitive advantage, as economists like to say). Likewise, Trek and Ridley have developed a reputation for quality, and they have neither the skills nor the factories

TABLE 1.1 MARKETS FOR BICYCLES

TYPE OF BICYCLE	COMPANIES AND PRICES (2011)
Mass Market Bicycles: Sold by mass merchandisers such as Target, Wal-Mart, Kmart, and Sears.	Huffy: \$90—\$140 Schwinn: \$140—\$240 Mantis: \$129—\$140 Mongoose: \$120—\$280
Dealer Bicycles: Sold by bicycle dealers — stores that sell only (or mostly) bicycles and bicycle equipment.	Trek: \$400—\$2500 Cannondale: \$500—\$2000 Giant: \$500—\$2500 Gary Fisher: \$600—\$2000 Mongoose: \$700—\$2000 Ridley: \$1300—\$2500 Scott: \$1000—\$3000 Ibis: \$2000 and up



to produce \$100 bicycles. Mongoose, on the other hand, straddles both markets. They produce mass market bicycles costing as little as \$120, but also high-quality dealer bicycles costing \$700 to \$2000.

After you buy your bike, you will need to lock it up carefully due to the unfortunate reality of yet another market—the black market for used bikes and their parts. We hope that you—and your bike—stay out of that market!

1.3 Real versus Nominal Prices

We often want to compare the price of a good today with what it was in the past or is likely to be in the future. To make such a comparison meaningful, we need to measure prices relative to an *overall price level*. In absolute terms, the price of a dozen eggs is many times higher today than it was 50 years ago. Relative to prices overall, however, it is actually lower. Therefore, we must be careful to correct for inflation when comparing prices across time. This means measuring prices in *real* rather than *nominal* terms.

- **nominal price** Absolute price of a good, unadjusted for inflation.

- **real price** Price of a good relative to an aggregate measure of prices; price adjusted for inflation.

- **Consumer Price Index** Measure of the aggregate price level.

- **Producer Price Index** Measure of the aggregate price level for intermediate products and wholesale goods.

The **nominal price** of a good (sometimes called its “current-dollar” price) is its absolute price. For example, the nominal price of a pound of butter was about \$0.87 in 1970, \$1.88 in 1980, about \$1.99 in 1990, and about \$3.42 in 2010. These are the prices you would have seen in supermarkets in those years. The **real price** of a good (sometimes called its “constant-dollar” price) is the price relative to an aggregate measure of prices. In other words, it is the price adjusted for inflation.

For consumer goods, the aggregate measure of prices most often used is the **Consumer Price Index (CPI)**. The CPI is calculated by the U.S. Bureau of Labor Statistics by surveying retail prices, and is published monthly. It records how the cost of a large market basket of goods purchased by a “typical” consumer changes over time. Percentage changes in the CPI measure the rate of inflation in the economy.

Sometimes we are interested in the prices of raw materials and other intermediate products bought by firms, as well as in finished products sold at wholesale to retail stores. In this case, the aggregate measure of prices often used is the **Producer Price Index (PPI)**. The PPI is also calculated by the U.S. Bureau of Labor Statistics and published monthly, and records how, on average, prices at the wholesale level change over time. Percentage changes in the PPI measure cost inflation and predict future changes in the CPI.

So which price index should you use to convert nominal prices to real prices? It depends on the type of product you are examining. If it is a product or service normally purchased by consumers, use the CPI. If instead it is a product normally purchased by businesses, use the PPI.

Because we are examining the price of butter in supermarkets, the relevant price index is the CPI. After correcting for inflation, do we find that the price of butter was more expensive in 2010 than in 1970? To find out, let’s calculate the 2010 price of butter in terms of 1970 dollars. The CPI was 38.8 in 1970 and rose to about 218.1 in 2010. (There was considerable inflation in the United States during the 1970s and early 1980s.) In 1970 dollars, the price of butter was

$$\frac{38.8}{218.1} \times \$3.42 = \$0.61$$



In real terms, therefore, the price of butter was lower in 2010 than it was in 1970.⁵ Put another way, the nominal price of butter went up by about 293 percent, while the CPI went up 462 percent. Relative to the aggregate price level, butter prices fell.

In this book, we will usually be concerned with real rather than nominal prices because consumer choices involve analyses of price comparisons. These relative prices can most easily be evaluated if there is a common basis of comparison. Stating all prices in real terms achieves this objective. Thus, even though we will often measure prices in dollars, we will be thinking in terms of the real purchasing power of those dollars.

EXAMPLE 1.3 THE PRICE OF EGGS AND THE PRICE OF A COLLEGE EDUCATION

In 1970, Grade A large eggs cost about 61 cents a dozen. In the same year, the average annual cost of a college education at a private four-year college, including room and board, was about \$2112. By 2010, the price of eggs had risen to \$1.54 a dozen, and the average cost of a college education was \$21,550. In real terms, were eggs more expensive in 2010 than in 1970? Had a college education become more expensive?

Table 1.2 shows the nominal price of eggs, the nominal cost of a college education, and the CPI for 1970–2010. (The CPI is based on 1983 = 100.)

TABLE 1.2 THE REAL PRICES OF EGGS AND OF A COLLEGE EDUCATION⁶

	1970	1980	1990	2000	2010
Consumer Price Index	38.8	82.4	130.7	172.2	218.1
Nominal Prices					
Grade A Large Eggs	\$0.61	\$0.84	\$1.01	\$0.91	\$1.54
College Education	\$2,112	\$3,502	\$7,619	\$12,976	\$21,550
Real Prices (\$1970)					
Grade A Large Eggs	\$0.61	\$0.40	\$0.30	\$0.21	\$0.27
College Education	\$2,112	\$1,649	\$2,262	\$2,924	\$3,835

⁵Two good sources of data on the national economy are the *Economic Report of the President* and the *Statistical Abstract of the United States*. Both are published annually and are available from the U.S. Government Printing Office.

⁶You can get data on the cost of a college education by visiting the National Center for Education Statistics and download the Digest of Education Statistics at <http://nces.ed.gov>. Historical and current data on the average retail price of eggs can be obtained from the Bureau of Labor Statistics (BLS) at <http://www.bls.gov>, by selecting CPI—Average Price Data.



Also shown are the *real* prices of eggs and college education in 1970 dollars, calculated as follows:

$$\text{Real price of eggs in 1980} = \frac{\text{CPI}_{1970}}{\text{CPI}_{1980}} \times \text{nominal price in 1980}$$

$$\text{Real price of eggs in 1990} = \frac{\text{CPI}_{1970}}{\text{CPI}_{1990}} \times \text{nominal price in 1990}$$

and so forth.

The table shows clearly that the real cost of a college education rose (by 82 percent) during this period, while the real cost of eggs fell (by 55 percent). It is these relative changes in prices that are important for the choices that consumers make, not the fact that both eggs and college cost more in nominal dollars today than they did in 1970.

In the table, we calculated real prices in terms of 1970 dollars, but we could just as easily have calculated them in terms of dollars of some other base year. For example, suppose we want to calculate the real price of eggs in 1990 *dollars*. Then:

$$\begin{aligned}\text{Real price of eggs in 1970} &= \frac{\text{CPI}_{1990}}{\text{CPI}_{1970}} \times \text{nominal price in 1970} \\ &= \frac{130.7}{38.8} \times 0.61 = 2.05\end{aligned}$$

$$\begin{aligned}\text{Real price of eggs in 2010} &= \frac{\text{CPI}_{1990}}{\text{CPI}_{2010}} \times \text{nominal price in 2010} \\ &= \frac{130.7}{218.1} \times 1.54 = 0.92\end{aligned}$$

$$\begin{aligned}\text{Percentage change in real price} &= \frac{\text{real price in 2010} - \text{real price in 1970}}{\text{real price in 1970}} \\ &= \frac{0.92 - 2.05}{2.05} = -0.55\end{aligned}$$

Notice that the percentage decline in real price is the same whether we use 1970 dollars or 1990 dollars as the base year.



EXAMPLE 1.4 THE MINIMUM WAGE

The federal minimum wage—first instituted in 1938 at a level of 25 cents per hour—has been increased periodically over the years. From 1991 through 1995, for example, it was \$4.25 an hour. Congress voted to raise it to \$4.75 in 1996 and then to \$5.15 in 1997. Legislation in 2007 to increase the minimum wage yet again would raise it to \$6.55 an hour in 2008 and \$7.25 in 2009.⁷

Figure 1.1 shows the minimum wage from 1938 through 2015, both in nominal terms and in 2000 constant dollars. Note that although the legislated minimum wage has steadily increased, in real terms the minimum wage today is not much different from what it was in the 1950s.

Nonetheless, the 2007 decision to increase the minimum wage was a difficult one. Although the higher minimum wage would provide a better standard of living for those workers who had been paid below the minimum, some analysts feared that it would also lead to increased unemployment among young and unskilled workers. The decision to increase the minimum wage, therefore, raises both normative and positive issues. The normative issue is whether any loss of teenage and low-skilled jobs is outweighed by two factors: (1) the direct benefits to those workers who now earn more as a result; and (2) any indirect benefits to other workers whose wages might be increased along

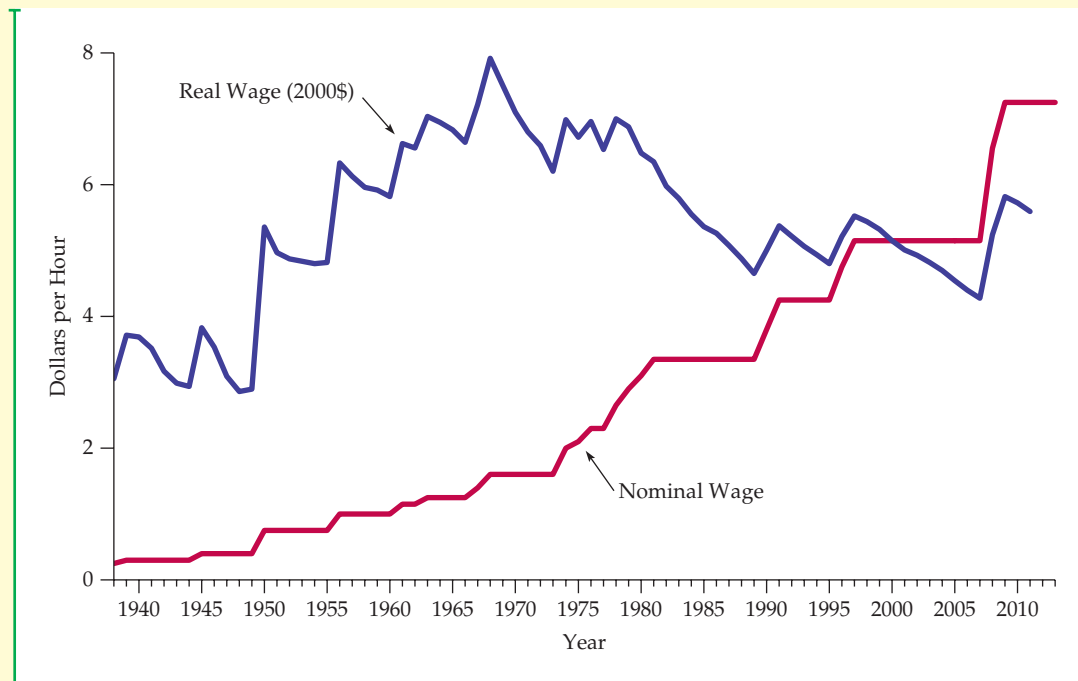


FIGURE 1.1
THE MINIMUM WAGE

In nominal terms, the minimum wage has increased steadily over the past 70 years. However, in real terms its 2010 level is below that of the 1970s.

⁷Some states also have minimum wages that are higher than the federal minimum wage. For example, in 2011 the minimum wage in Massachusetts was \$8.00 per hour, in New York it was \$7.25, and in California it was \$8.00 and scheduled to increase to \$8.00 in 2008. You can learn more about the minimum wage at <http://www.dol.gov>.



with the wages of those at the bottom of the pay scale.

An important positive issue is how many fewer workers (if any) would be able to get jobs with a higher minimum wage. As we will see in Chapter 14, this issue is still hotly debated. Statistical studies have suggested that an increase in the minimum

wage of about 10 percent would increase teenage unemployment by 1 to 2 percent. (The actual increase from \$5.15 to \$7.25 represents a 41-percent increase.) However, one review of the evidence questions whether there are any significant unemployment effects.⁸

1.4 Why Study Microeconomics?

We think that after reading this book you will have no doubt about the importance and broad applicability of microeconomics. In fact, one of our major goals is to show you how to apply microeconomic principles to actual decision-making problems. Nonetheless, some extra motivation early on never hurts. Here are two examples that not only show the use of microeconomics in practice, but also provide a preview of this book.

Corporate Decision Making: The Toyota Prius

In 1997, Toyota Motor Corporation introduced the Prius in Japan, and started selling it worldwide in 2001. The Prius, the first hybrid car to be sold in the United States, can run off both a gasoline engine and a battery, and the momentum of the car charges the battery. Hybrid cars are more energy efficient than cars with just a gasoline engine; the Prius, for example, can get 45 to 55 miles per gallon. The Prius was a big success, and within a few years other manufacturers began introducing hybrid versions of some of their cars.

The design and efficient production of the Prius involved not only some impressive engineering, but a lot of economics as well. First, Toyota had to think carefully about how the public would react to the design and performance of this new product. How strong would demand be initially, and how fast would it grow? How would demand depend on the prices that Toyota charged? Understanding consumer preferences and trade-offs and predicting demand and its responsiveness to price are essential to Toyota and every other automobile manufacturer. (We discuss consumer preferences and demand in Chapters 3, 4, and 5.)

Next, Toyota had to be concerned with the cost of manufacturing these cars — whether produced in Japan or, starting in 2010, in the United States. How high would production costs be? How would the cost of each car depend on the total number of cars produced each year? How would the cost of labor and the prices of steel and other raw materials affect costs? How much and how fast would costs decline as managers and workers gained experience with the production

⁸The first study is David Neumark and William Wascher, “Employment Effects of Minimum and Subminimum Wages: Panel Data on State Minimum Wage Laws,” *Industrial and Labor Relations Review* 46 (October 1992): 55–81. A review of the literature appears in David Card and Alan Krueger, *Myth and Measurement: The New Economics of the Minimum Wage* (Princeton: Princeton University Press, 1995).



process? And to maximize profits, how many of these cars should Toyota plan to produce each year? (We discuss production and cost in Chapters 6 and 7, and the profit-maximizing choice of output in Chapters 8 and 10.)

Toyota also had to design a pricing strategy and consider how competitors would react to it. Although the Prius was the first hybrid car, Toyota knew that it would compete with other small fuel-efficient cars, and that soon other manufacturers would introduce their own hybrid cars. Should Toyota charge a relatively low price for a basic stripped-down version of the Prius and high prices for individual options like leather seats? Or would it be more profitable to make these options “standard” items and charge a higher price for the whole package? Whatever pricing strategy Toyota chose, how were competitors likely to react? Would Ford or Nissan try to undercut by lowering the prices of its smaller cars, or rush to bring out their own hybrid cars at lower prices? Might Toyota be able to deter Ford and Nissan from lowering prices by threatening to respond with its own price cuts? (We discuss pricing in Chapters 10 and 11, and competitive strategy in Chapters 12 and 13.)

Manufacturing the Prius required large investments in new capital equipment, so Toyota had to consider both the risks and possible outcomes of its decisions. Some of this risk was due to uncertainty over the future price of oil and thus the price of gasoline (lower gasoline prices would reduce the demand for small fuel-efficient cars). Some of the risk was due to uncertainty over the wages that Toyota would have to pay its workers at its plants in Japan and in the United States. (Oil and other commodity markets are discussed in Chapters 2 and 9. Labor markets and the impact of unions are discussed in Chapter 14. Investment decisions and the implications of uncertainty are discussed in Chapters 5 and 15.)

Toyota also had to worry about organizational problems. Toyota is an integrated firm in which separate divisions produce engines and parts and then assemble finished cars. How should the managers of different divisions be rewarded? What price should the assembly division be charged for the engines it receives from another division? (We discuss internal pricing and organizational incentives for the integrated firm in Chapters 11 and 17.)

Finally, Toyota had to think about its relationship to the government and the effects of regulatory policies. For example, all of its cars sold in the United States must meet federal emissions standards, and U.S. production-line operations must comply with health and safety regulations. How might those regulations and standards change over time? How would they affect costs and profits? (We discuss the role of government in limiting pollution and promoting health and safety in Chapter 18.)

Public Policy Design: Fuel Efficiency Standards for the Twenty-First Century

In 1975, the U.S. government imposed regulations designed to improve the average fuel economy of domestically-sold cars and light trucks (including vans and sport utility vehicles). The CAFE (Corporate Average Fuel Economy) standards have become increasingly stringent over the years. In 2007, President George W. Bush signed into law the Energy Independence and Security Act, which required automakers to boost fleet wide gas mileage to 35 miles per gallon (mpg) by 2020. In 2011, the Obama administration pushed the 35 mpg target forward to 2016, and (with the agreement of 13 auto companies) set a standard of 55 mpg for 2020. While the program’s primary goal is to increase



energy security by reducing the U.S. dependence on imported oil, it would also generate substantial environmental benefits, such as a reduction in greenhouse gas emissions.

A number of important decisions have to be made when designing a fuel efficiency program, and most of those decisions involve economics. First, the government must evaluate the monetary impact of the program on consumers. Higher fuel economy standards will increase the cost of purchasing a car (the cost of achieving higher fuel economy will be borne in part by consumers), but will lower the cost of operating it (gas mileage will be higher). Analyzing the ultimate impact on consumers means analyzing consumer preferences and demand. For example, would consumers drive less and spend more of their income on other goods? If so, would they be nearly as well off? (Consumer preferences and demand are discussed in Chapters 3 and 4).

Before imposing CAFE standards, it is important to estimate the likely impact those standards will have on the cost of producing cars and light trucks. Might automobile companies minimize cost increases by using new lightweight materials or by changing the footprint of new model cars? (Production and cost are discussed in Chapters 6 and 7.) Then the government needs to know how changes in production costs will affect the production levels and prices of new automobiles and light trucks. Are the additional costs likely to be absorbed by manufacturers or passed on to consumers in the form of higher prices? (Output determination is discussed in Chapter 8 and pricing in Chapters 10 through 13.)

The government must also ask why problems related to oil consumption are not solved by our market-oriented economy. One answer is that oil prices are determined in part by a cartel (OPEC) that is able to push the price of oil above competitive levels. (Pricing in markets in which firms have the power to control prices are discussed in Chapters 10 through 12.) Finally, the high U.S. demand for oil has led to a substantial outflow of dollars to the oil-producing countries, which in turn has created political and security issues that go beyond the confines of economics. What economics can do, however, is help us evaluate how best to reduce our dependence on foreign oil. Are standards like those of the CAFE program preferred to fees on oil consumption? What are the environmental implications of increasingly stringent standards? (These problems are discussed in Chapter 18.)

These are just two examples of how microeconomics can be applied in the arenas of private and public-policy decision making. You will discover many more applications as you read this book.

SUMMARY

1. Microeconomics is concerned with the decisions made by individual economic units—consumers, workers, investors, owners of resources, and business firms. It is also concerned with the interaction of consumers and firms to form markets and industries.
2. Microeconomics relies heavily on the use of theory, which can (by simplification) help to explain how economic units behave and to predict what behavior will occur in the future. Models are mathematical representations of theories that can help in this explanation and prediction process.
3. Microeconomics is concerned with positive questions that have to do with the explanation and prediction of phenomena. But microeconomics is also important for normative analysis, in which we ask what choices are best—for a firm or for society as a whole. Normative analyses must often be combined with individual value judgments because issues of equity and fairness as well as of economic efficiency may be involved.
4. A *market* refers to a collection of buyers and sellers who interact, and to the possibility for sales and purchases that result from that interaction.



Microeconomics involves the study of both perfectly competitive markets, in which no single buyer or seller has an impact on price, and noncompetitive markets, in which individual entities can affect price.

5. The market price is established by the interaction of buyers and sellers. In a perfectly competitive market, a single price will usually prevail. In markets that are not perfectly competitive, different sellers might charge different prices. In this case, the market price refers to the average prevailing price.

6. When discussing a market, we must be clear about its extent in terms of both its geographic boundaries and the range of products to be included in it. Some markets (e.g., housing) are highly localized, whereas others (e.g., gold) are global in nature.
7. To account for the effects of inflation, we measure real (or constant-dollar) prices, rather than nominal (or current-dollar) prices. Real prices use an aggregate price index, such as the CPI, to correct for inflation.

QUESTIONS FOR REVIEW

1. It is often said that a good theory is one that can be refuted by an empirical, data-oriented study. Explain why a theory that cannot be evaluated empirically is not a good theory.
2. Which of the following two statements involves positive economic analysis and which normative? How do the two kinds of analysis differ?
 - a. Gasoline rationing (allocating to each individual a maximum amount of gasoline that can be purchased each year) is poor social policy because it interferes with the workings of the competitive market system.
 - b. Gasoline rationing is a policy under which more people are made worse off than are made better off.
3. Suppose the price of regular-octane gasoline were 20 cents per gallon higher in New Jersey than in Oklahoma. Do you think there would be an opportunity for arbitrage (i.e., that firms could buy gas in Oklahoma and then sell it at a profit in New Jersey)? Why or why not?
4. In Example 1.3, what economic forces explain why the real price of eggs has fallen while the real price of a college education has increased? How have these changes affected consumer choices?
5. Suppose that the Japanese yen rises against the U.S. dollar—that is, it will take more dollars to buy a given amount of Japanese yen. Explain why this increase simultaneously increases the real price of Japanese cars for U.S. consumers and lowers the real price of U.S. automobiles for Japanese consumers.
6. The price of long-distance telephone service fell from 40 cents per minute in 1996 to 22 cents per minute in 1999, a 45-percent (18 cents/40 cents) decrease. The Consumer Price Index increased by 10 percent over this period. What happened to the real price of telephone service?

EXERCISES

1. Decide whether each of the following statements is true or false and explain why:
 - a. Fast-food chains like McDonald's, Burger King, and Wendy's operate all over the United States. Therefore, the market for fast food is a national market.
 - b. People generally buy clothing in the city in which they live. Therefore, there is a clothing market in, say, Atlanta that is distinct from the clothing market in Los Angeles.
 - c. Some consumers strongly prefer Pepsi and some strongly prefer Coke. Therefore, there is no single market for colas.
2. The following table shows the average retail price of butter and the Consumer Price Index from 1980 to 2010, scaled so that the CPI = 100 in 1980.

	1980	1990	2000	2010
CPI	100	158.56	208.98	218.06
Retail price of butter (salted, grade AA, per lb.)	\$1.88	\$1.99	\$2.52	\$2.88

- a. Calculate the real price of butter in 1980 dollars. Has the real price increased/decreased/stayed the same from 1980 to 2000? From 1980 to 2010?
- b. What is the percentage change in the real price (1980 dollars) from 1980 to 2000? From 1980 to 2010?
- c. Convert the CPI into 1990 = 100 and determine the real price of butter in 1990 dollars.



- d. What is the percentage change in real price (1990 dollars) from 1980 to 2000? Compare this with your answer in (b). What do you notice? Explain.
3. At the time this book went to print, the minimum wage was \$7.25. To find the current value of the CPI, go to <http://www.bls.gov/cpi/home.htm>. Click on "CPI Tables," which is found on the left side of the web page. Then, click on "Table Containing History of CPI-U U.S. All Items Indexes and Annual Percent Changes from 1913 to Present." This will give you the CPI from 1913 to the present.
 - a. With these values, calculate the current real minimum wage in 1990 dollars.
 - b. Stated in real 1990 dollars, what is the percentage change in the real minimum wage from 1985 to the present?

CHAPTER 2

The Basics of Supply and Demand

One of the best ways to appreciate the relevance of economics is to begin with the basics of supply and demand. Supply-demand analysis is a fundamental and powerful tool that can be applied to a wide variety of interesting and important problems. To name a few:

- Understanding and predicting how changing world economic conditions affect market price and production
- Evaluating the impact of government price controls, minimum wages, price supports, and production incentives
- Determining how taxes, subsidies, tariffs, and import quotas affect consumers and producers

We begin with a review of how supply and demand curves are used to describe the *market mechanism*. Without government intervention (e.g., through the imposition of price controls or some other regulatory policy), supply and demand will come into equilibrium to determine both the market price of a good and the total quantity produced. What that price and quantity will be depends on the particular characteristics of supply and demand. Variations of price and quantity over time depend on the ways in which supply and demand respond to other economic variables, such as aggregate economic activity and labor costs, which are themselves changing.

We will, therefore, discuss the characteristics of supply and demand and show how those characteristics may differ from one market to another. Then we can begin to use supply and demand curves to understand a variety of phenomena—for example, why the prices of some basic commodities have fallen steadily over a long period while the prices of others have experienced sharp fluctuations; why shortages occur in certain markets; and why announcements about plans for future government policies or predictions about future economic conditions can affect markets well before those policies or conditions become reality.

Besides understanding *qualitatively* how market price and quantity are determined and how they can vary over time, it is also important to learn how they can be analyzed *quantitatively*. We will see how simple “back of the envelope” calculations can be used to analyze and predict evolving market conditions. We will also show how markets respond



CHAPTER OUTLINE

2.1	Supply and Demand	22
2.2	The Market Mechanism	25
2.3	Changes in Market Equilibrium	26
2.4	Elasticities of Supply and Demand	33
2.5	Short-Run versus Long-Run Elasticities	39
*2.6	Understanding and Predicting the Effects of Changing Market Conditions	48
2.7	Effects of Government Intervention—Price Controls	58

LIST OF EXAMPLES

2.1	The Price of Eggs and the Price of a College Education Revisited	28
2.2	Wage Inequality in the United States	29
2.3	The Long-Run Behavior of Natural Resource Prices	29
2.4	The Effects of 9/11 on the Supply and Demand for New York City Office Space	31
2.5	The Market for Wheat	37
2.6	The Demand for Gasoline and Automobiles	43
2.7	The Weather in Brazil and the Price of Coffee in New York	46
2.8	The Behavior of Copper Prices	52
2.9	Upheaval in the World Oil Market	54
2.10	Price Controls and Natural Gas Shortages	59



both to domestic and international macroeconomic fluctuations and to the effects of government interventions. We will try to convey this understanding through simple examples and by urging you to work through some exercises at the end of the chapter.

2.1 Supply and Demand

The basic model of supply and demand is the workhorse of microeconomics. It helps us understand why and how prices change, and what happens when the government intervenes in a market. The supply-demand model combines two important concepts: a *supply curve* and a *demand curve*. It is important to understand precisely what these curves represent.

• **supply curve** Relationship between the quantity of a good that producers are willing to sell and the price of the good.

The Supply Curve

The **supply curve** shows the quantity of a good that producers are willing to sell at a given price, holding constant any other factors that might affect the quantity supplied. The curve labeled S in Figure 2.1 illustrates this. The vertical axis of the graph shows the price of a good, P , measured in dollars per unit. This is the price that sellers receive for a given quantity supplied. The horizontal axis shows the total quantity supplied, Q , measured in the number of units per period.

The supply curve is thus a relationship between the quantity supplied and the price. We can write this relationship as an equation:

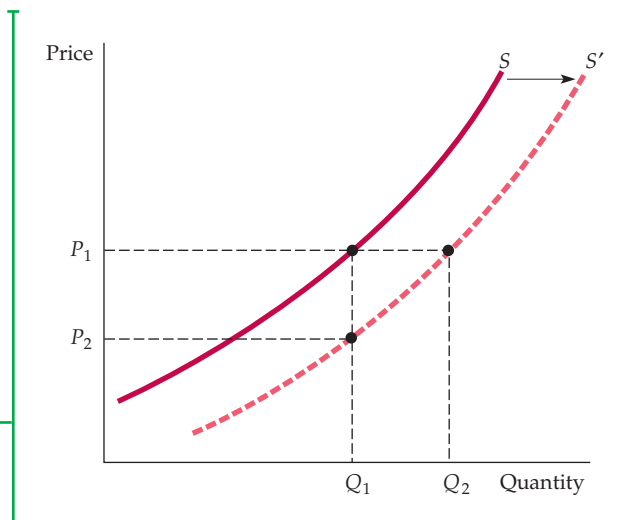
$$Q_S = Q_S(P)$$

Or we can draw it graphically, as we have done in Figure 2.1.

Note that the supply curve in Figure 2.1 slopes upward. In other words, the higher the price, the more that firms are able and willing to produce and sell. For example, a higher price may enable current firms to expand production by hiring extra workers or by having existing workers work overtime (at greater cost to the firm). Likewise, they may expand production over a longer period of time by increasing the size of their plants. A higher price may also attract new

FIGURE 2.1
THE SUPPLY CURVE

The supply curve, labeled S in the figure, shows how the quantity of a good offered for sale changes as the price of the good changes. The supply curve is upward sloping: The higher the price, the more firms are able and willing to produce and sell. If production costs fall, firms can produce the same quantity at a lower price or a larger quantity at the same price. The supply curve then shifts to the right (from S to S').





firms to the market. These newcomers face higher costs because of their inexperience in the market and would therefore have found entry uneconomical at a lower price.

OTHER VARIABLES THAT AFFECT SUPPLY The quantity supplied can depend on other variables besides price. For example, the quantity that producers are willing to sell depends not only on the price they receive but also on their production costs, including wages, interest charges, and the costs of raw materials. The supply curve labeled S in Figure 2.1 was drawn for particular values of these other variables. A change in the values of one or more of these variables translates into a shift in the supply curve. Let's see how this might happen.

The supply curve S in Figure 2.1 says that at a price P_1 , the quantity produced and sold would be Q_1 . Now suppose that the cost of raw materials *falls*. How does this affect the supply curve?

Lower raw material costs—indeed, lower costs of any kind—make production more profitable, encouraging existing firms to expand production and enabling new firms to enter the market. If at the same time the market price stayed constant at P_1 , we would expect to observe a greater quantity supplied. Figure 2.1 shows this as an increase from Q_1 to Q_2 . When production costs *decrease*, output *increases* no matter what the market price happens to be. *The entire supply curve thus shifts to the right*, which is shown in the figure as a shift from S to S' .

Another way of looking at the effect of lower raw material costs is to imagine that the quantity produced stays fixed at Q_1 and then ask what price firms would require to produce this quantity. Because their costs are lower, they would accept a lower price— P_2 . This would be the case no matter what quantity was produced. Again, we see in Figure 2.1 that the supply curve must shift to the right.

We have seen that the response of quantity supplied to changes in price can be represented by movements *along the supply curve*. However, the response of supply to changes in other supply-determining variables is shown graphically as a *shift of the supply curve itself*. To distinguish between these two graphical depictions of supply changes, economists often use the phrase *change in supply* to refer to shifts in the supply curve, while reserving the phrase *change in the quantity supplied* to apply to movements along the supply curve.

The Demand Curve

The **demand curve** shows how much of a good consumers are willing to buy as the price per unit changes. We can write this relationship between quantity demanded and price as an equation:

$$Q_D = Q_D(P)$$

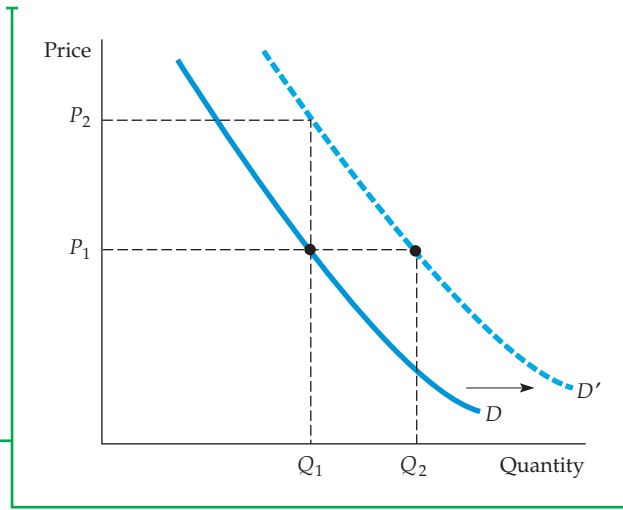
or we can draw it graphically, as in Figure 2.2. Note that the demand curve in that figure, labeled D , slopes *downward*: Consumers are usually ready to buy more if the price is lower. For example, a lower price may encourage consumers who have already been buying the good to consume larger quantities. Likewise, it may allow other consumers who were previously unable to afford the good to begin buying it.

Of course the quantity of a good that consumers are willing to buy can depend on other things besides its price. *Income* is especially important. With greater incomes, consumers can spend more money on any good, and some consumers will do so for most goods.

• **demand curve** Relationship between the quantity of a good that consumers are willing to buy and the price of the good.

FIGURE 2.2 THE DEMAND CURVE

The demand curve, labeled D , shows how the quantity of a good demanded by consumers depends on its price. The demand curve is downward sloping; holding other things equal, consumers will want to purchase more of a good as its price goes down. The quantity demanded may also depend on other variables, such as income, the weather, and the prices of other goods. For most products, the quantity demanded increases when income rises. A higher income level shifts the demand curve to the right (from D to D').



SHIFTING THE DEMAND CURVE Let's see what happens to the demand curve if income levels increase. As you can see in Figure 2.2, if the market price were held constant at P_1 , we would expect to see an increase in the quantity demanded—say, from Q_1 to Q_2 , as a result of consumers' higher incomes. Because this increase would occur no matter what the market price, the result would be a *shift to the right of the entire demand curve*. In the figure, this is shown as a shift from D to D' . Alternatively, we can ask what price consumers would pay to purchase a given quantity Q_1 . With greater income, they should be willing to pay a higher price—say, P_2 instead of P_1 in Figure 2.2. Again, *the demand curve will shift to the right*. As we did with supply, we will use the phrase *change in demand* to refer to shifts in the demand curve, and reserve the phrase *change in the quantity demanded* to apply to movements along the demand curve.¹

• **substitutes** Two goods for which an increase in the price of one leads to an increase in the quantity demanded of the other.

• **complements** Two goods for which an increase in the price of one leads to a decrease in the quantity demanded of the other.

SUBSTITUTE AND COMPLEMENTARY GOODS Changes in the prices of related goods also affect demand. Goods are **substitutes** when an increase in the price of one leads to an increase in the quantity demanded of the other. For example, copper and aluminum are substitute goods. Because one can often be substituted for the other in industrial use, *the quantity of copper demanded will increase if the price of aluminum increases*. Likewise, beef and chicken are substitute goods because most consumers are willing to shift their purchases from one to the other when prices change.

Goods are **complements** when an increase in the price of one leads to a decrease in the quantity demanded of the other. For example, automobiles and gasoline are complementary goods. Because they tend to be used together, a decrease in the price of gasoline increases the quantity demanded for automobiles. Likewise, computers and computer software are complementary goods. The price of computers has dropped dramatically over the past decade, fueling an increase not only in purchases of computers, but also purchases of software packages.

We attributed the shift to the right of the demand curve in Figure 2.2 to an increase in income. However, this shift could also have resulted from either an increase in the price of a substitute good or a decrease in the price of a

¹Mathematically, we can write the demand curve as

$$Q_D = D(P, I)$$

where I is disposable income. When we draw a demand curve, we are keeping I fixed.



complementary good. Or it might have resulted from a change in some other variable, such as the weather. For example, demand curves for skis and snowboards will shift to the right when there are heavy snowfalls.

2.2 The Market Mechanism

The next step is to put the supply curve and the demand curve together. We have done this in Figure 2.3. The vertical axis shows the price of a good, P , again measured in dollars per unit. This is now the price that sellers receive for a given quantity supplied, and the price that buyers will pay for a given quantity demanded. The horizontal axis shows the total quantity demanded and supplied, Q , measured in number of units per period.

EQUILIBRIUM The two curves intersect at the **equilibrium**, or **market-clearing, price** and quantity. At this price (P_0 in Figure 2.3), the quantity supplied and the quantity demanded are just equal (to Q_0). The **market mechanism** is the tendency in a free market for the price to change until the market *clears*—i.e., until the quantity supplied and the quantity demanded are equal. At this point, because there is neither excess demand nor excess supply, there is no pressure for the price to change further. Supply and demand might not always be in equilibrium, and some markets might not clear quickly when conditions change suddenly. The *tendency*, however, is for markets to clear.

To understand why markets tend to clear, suppose the price were initially above the market-clearing level—say, P_1 in Figure 2.3. Producers will try to produce and sell more than consumers are willing to buy. A **surplus**—a situation in which the quantity supplied exceeds the quantity demanded—will result. To sell this surplus—or at least to prevent it from growing—producers would begin to lower prices. Eventually, as price fell, quantity demanded would increase, and quantity supplied would decrease until the equilibrium price P_0 was reached.

The opposite would happen if the price were initially below P_0 —say, at P_2 . A **shortage**—a situation in which the quantity demanded exceeds the quantity

- **equilibrium (or market-clearing) price** Price that equates the quantity supplied to the quantity demanded.

- **market mechanism** Tendency in a free market for price to change until the market clears.

- **surplus** Situation in which the quantity supplied exceeds the quantity demanded.

- **shortage** Situation in which the quantity demanded exceeds the quantity supplied.

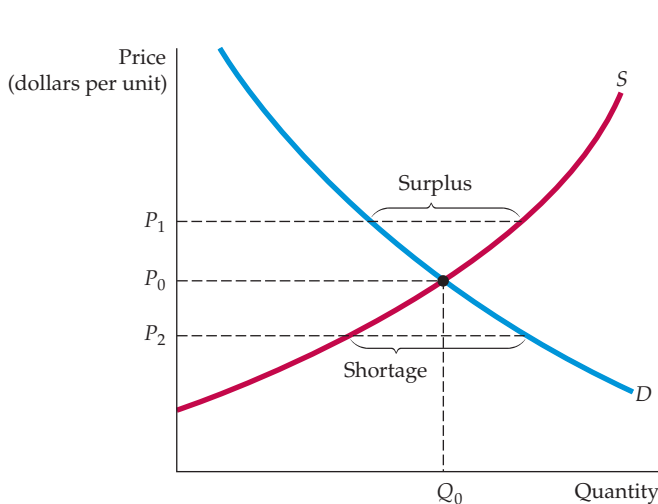


FIGURE 2.3
SUPPLY AND DEMAND

The market clears at price P_0 and quantity Q_0 . At the higher price P_1 , a surplus develops, so price falls. At the lower price P_2 , there is a shortage, so price is bid up.



supplied—would develop, and consumers would be unable to purchase all they would like. This would put upward pressure on price as consumers tried to outbid one another for existing supplies and producers reacted by increasing price and expanding output. Again, the price would eventually reach P_0 .

WHEN CAN WE USE THE SUPPLY-DEMAND MODEL? When we draw and use supply and demand curves, we are assuming that at any given price, a given quantity will be produced and sold. This assumption makes sense only if a market is at least roughly *competitive*. By this we mean that both sellers and buyers should have little *market power*—i.e., little ability *individually* to affect the market price.

Suppose instead that supply were controlled by a single producer—a monopolist. In this case, there will no longer be a simple one-to-one relationship between price and the quantity supplied. Why? Because a monopolist's behavior depends on the shape and position of the demand curve. If the demand curve shifts in a particular way, it may be in the monopolist's interest to keep the quantity fixed but change the price, or to keep the price fixed and change the quantity. (How this could occur is explained in Chapter 10.) Thus when we work with supply and demand curves, we implicitly assume that we are referring to a competitive market.

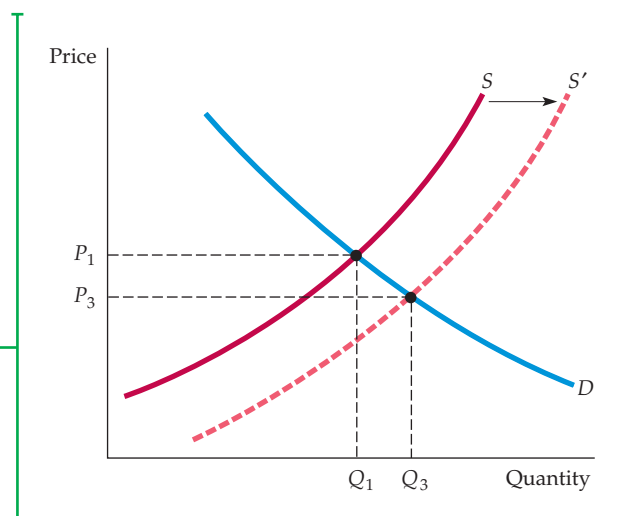
2.3 Changes in Market Equilibrium

We have seen how supply and demand curves shift in response to changes in such variables as wage rates, capital costs, and income. We have also seen how the market mechanism results in an equilibrium in which the quantity supplied equals the quantity demanded. Now we will see how that equilibrium changes in response to shifts in the supply and demand curves.

Let's begin with a shift in the supply curve. In Figure 2.4, the supply curve has shifted from S to S' (as it did in Figure 2.1), perhaps as a result of a decrease in the price of raw materials. As a result, the market price drops (from P_1 to P_3), and the total quantity produced increases (from Q_1 to Q_3). This is what we

FIGURE 2.4
NEW EQUILIBRIUM FOLLOWING SHIFT IN SUPPLY

When the supply curve shifts to the right, the market clears at a lower price P_3 and a larger quantity Q_3 .





would expect: Lower costs result in lower prices and increased sales. (Indeed, gradual decreases in costs resulting from technological progress and better management are an important driving force behind economic growth.)

Figure 2.5 shows what happens following a rightward shift in the demand curve resulting from, say, an increase in income. A new price and quantity result after demand comes into equilibrium with supply. As shown in Figure 2.5, we would expect to see consumers pay a higher price, P_3 , and firms produce a greater quantity, Q_3 , as a result of an increase in income.

In most markets, both the demand and supply curves shift from time to time. Consumers' disposable incomes change as the economy grows (or contracts, during economic recessions). The demands for some goods shift with the seasons (e.g., fuels, bathing suits, umbrellas), with changes in the prices of related goods (an increase in oil prices increases the demand for natural gas), or simply with changing tastes. Similarly, wage rates, capital costs, and the prices of raw materials also change from time to time, and these changes shift the supply curve.

Supply and demand curves can be used to trace the effects of these changes. In Figure 2.6, for example, shifts to the right of both supply and demand result in a slightly higher price (from P_1 to P_2) and a much larger quantity (from Q_1 to Q_2). In general, price and quantity will change depending both on how much the supply and demand curves shift and on the shapes of those curves. To predict the sizes and directions of such changes, we must be able to characterize quantitatively the dependence of supply and demand on price and other variables. We will turn to this task in the next section.

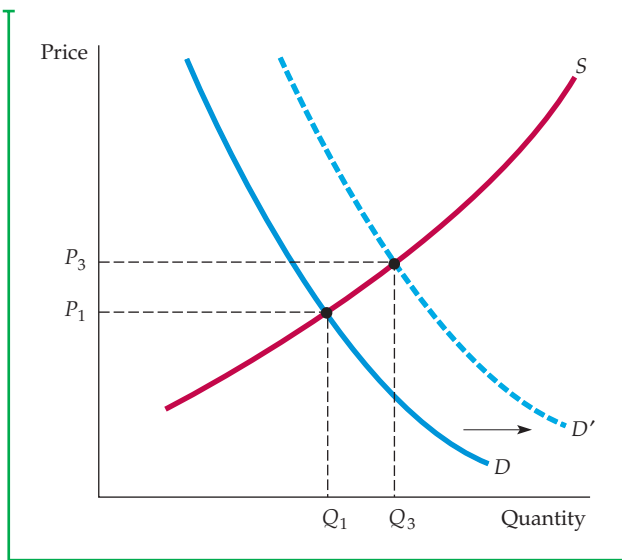


FIGURE 2.5
NEW EQUILIBRIUM FOLLOWING
SHIFT IN DEMAND

When the demand curve shifts to the right, the market clears at a higher price P_3 and a larger quantity Q_3 .

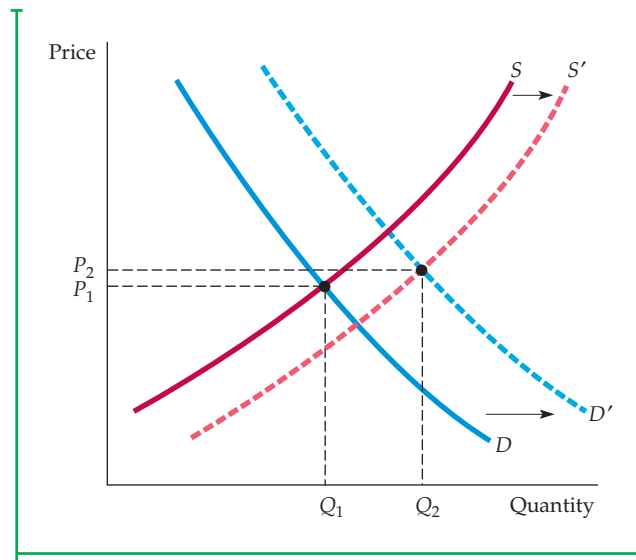


FIGURE 2.6
NEW EQUILIBRIUM FOLLOWING SHIFTS
IN SUPPLY AND DEMAND

Supply and demand curves shift over time as market conditions change. In this example, rightward shifts of the supply and demand curves lead to a slightly higher price and a much larger quantity. In general, changes in price and quantity depend on the amount by which each curve shifts and the shape of each curve.



EXAMPLE 2.1 THE PRICE OF EGGS AND THE PRICE OF A COLLEGE EDUCATION REVISITED

In Example 1.3 (page 13), we saw that from 1970 to 2010, the real (constant-dollar) price of eggs fell by 55 percent, while the real price of a college education rose by 82 percent. What caused this large decline in egg prices and large increase in the price of college?



eggs declined sharply while total annual consumption increased (from 5300 million dozen to 6392 million dozen).

As for college, supply and demand shifted in the opposite directions. Increases in the costs of equipping and maintaining modern classrooms, laboratories, and libraries, along with increases

We can understand these price changes by examining the behavior of supply and demand for each good, as shown in Figure 2.7. For eggs, the mechanization of poultry farms sharply reduced the cost of producing eggs, shifting the supply curve downward. At the same time, the demand curve for eggs shifted to the left as a more health-conscious population changed its eating habits and tended to avoid eggs. As a result, the real price of

in faculty salaries, pushed the supply curve up. At the same time, the demand curve shifted to the right as a larger percentage of a growing number of high school graduates decided that a college education was essential. Thus, despite the increase in price, 2010 found 12.5 million students enrolled in four-year undergraduate college degree programs, compared with 6.9 million in 1970.

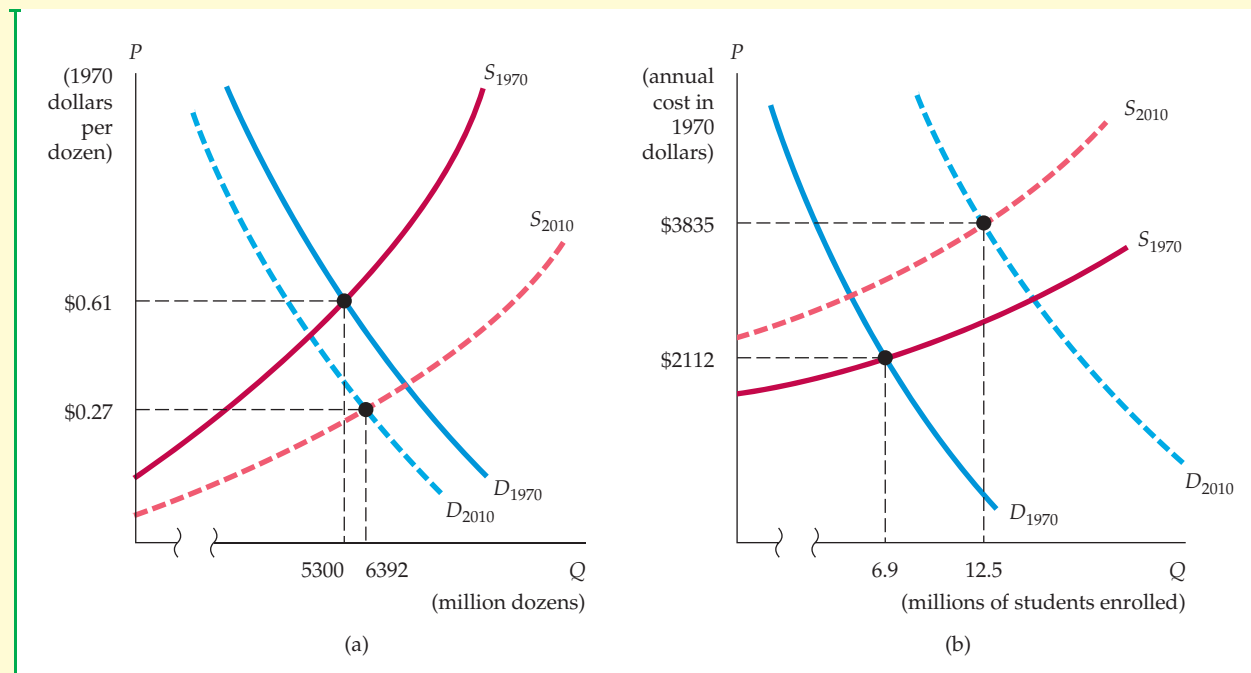


FIGURE 2.7
(a) MARKET FOR EGGS (b) MARKET FOR COLLEGE EDUCATION

(a) The supply curve for eggs shifted downward as production costs fell; the demand curve shifted to the left as consumer preferences changed. As a result, the real price of eggs fell sharply and egg consumption rose. (b) The supply curve for a college education shifted up as the costs of equipment, maintenance, and staffing rose. The demand curve shifted to the right as a growing number of high school graduates desired a college education. As a result, both price and enrollments rose sharply.



EXAMPLE 2.2 WAGE INEQUALITY IN THE UNITED STATES

Although the U.S. economy has grown vigorously over the past two decades, the gains from this growth have not been shared equally by all. The wages of skilled high-income workers have grown substantially, while the wages of unskilled low-income workers have, in real terms, actually fallen slightly. Overall, there has been growing inequality in the distribution of earnings, a phenomenon which began around 1980 and has accelerated in recent years. For example, from 1978 to 2009, people in the top 20 percent of the income distribution experienced an increase in their average real (inflation-adjusted) pretax household income of 45 percent, while those in the bottom 20 percent saw their average real pretax income increase by only 4 percent.²

Why has income distribution become so much more unequal during the past two decades? The answer is in the supply and demand for workers. While the supply of unskilled workers—people with limited educations—has grown substantially, the demand for them has risen only slightly. This shift of the supply curve to the right, combined with little movement of the demand curve, has caused wages of unskilled workers to fall. On the other hand,

while the supply of skilled workers—e.g., engineers, scientists, managers, and economists—has grown slowly, the demand has risen dramatically, pushing wages up. (We leave it to you as an exercise to draw supply and demand curves and show how they have shifted, as was done in Example 2.1.)

These trends are evident in the behavior of wages for different categories of employment. From 1980 to 2009, for example, the real (inflation-adjusted) weekly earnings of skilled workers (such as finance, insurance, and real estate workers) rose by more than 20 percent. Over the same period, the weekly real incomes of relatively unskilled workers (such as retail trade workers) rose by only 5 percent.³

Most projections point to a continuation of this phenomenon during the coming decade. As the high-tech sectors of the American economy grow, the demand for highly skilled workers is likely to increase further. At the same time, the computerization of offices and factories will further reduce the demand for unskilled workers. (This trend is discussed further in Example 14.7.) These changes can only exacerbate wage inequality.

EXAMPLE 2.3 THE LONG-RUN BEHAVIOR OF NATURAL RESOURCE PRICES

Many people are concerned about the earth's natural resources. At issue is whether our energy and mineral resources are likely to be depleted in the near future, leading to sharp price increases that could bring an end to economic growth. An analysis of supply and demand can give us some perspective.



The earth does indeed have only a finite amount of mineral resources, such as copper, iron, coal, and oil. During the past century, however, the prices of these and most other natural resources have declined or remained roughly constant relative to overall prices. Figure 2.8, for example,

²In *after-tax* terms, the growth of inequality has been even greater; the average real after-tax income of the bottom 20 percent of the distribution *fell* over this period. For historical data on income inequality in the United States, see the Historical Income Inequality Tables at the U.S. Census Bureau Web site: <http://www.census.gov/>.

³For detailed earnings data, visit the Detailed Statistics section of the web site of the Bureau of Labor Statistics (BLS): <http://www.bls.gov/>. Select Employment, Hours, and Earnings from the Current Employment Statistics survey (National).

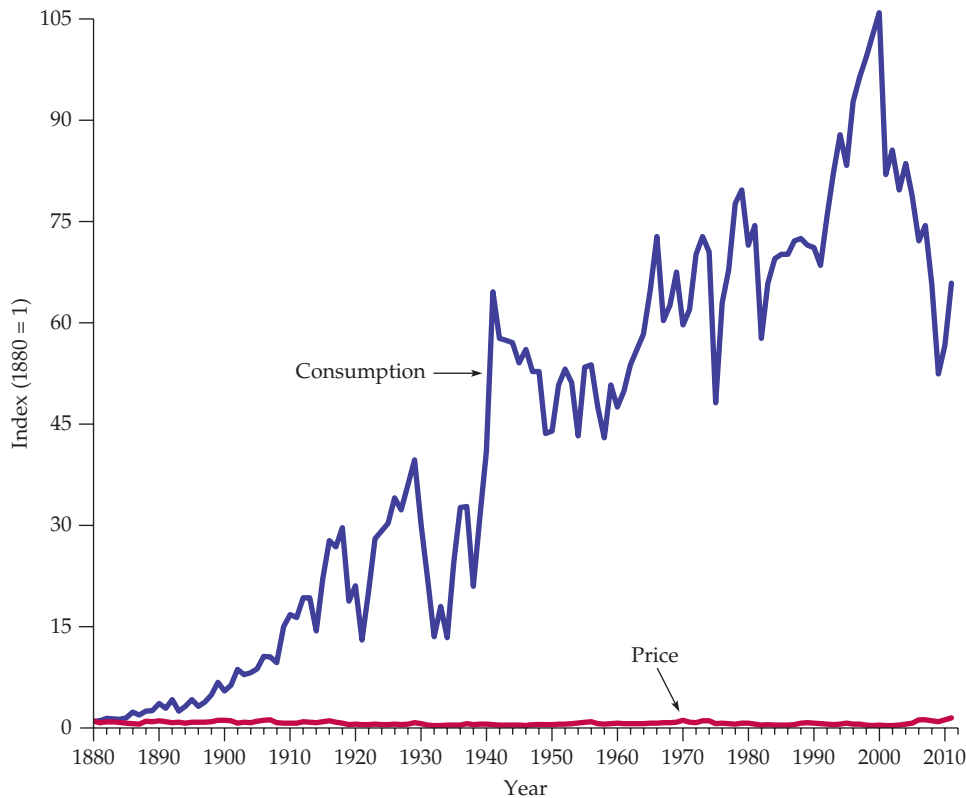


FIGURE 2.8
CONSUMPTION AND PRICE OF COPPER

Although annual consumption of copper has increased about a hundredfold, the real (inflation-adjusted) price has not changed much.

shows the price of copper in real terms (adjusted for inflation), together with the quantity consumed from 1880 to 2010. (Both are shown as an index, with 1880 = 1.) Despite short-term variations in price, no significant long-term increase has occurred, even though annual consumption is now about 100 times greater than in 1880. Similar patterns hold for other mineral resources, such as iron, oil, and coal.⁴

How can we explain this huge increase in copper consumption but very little change in price?

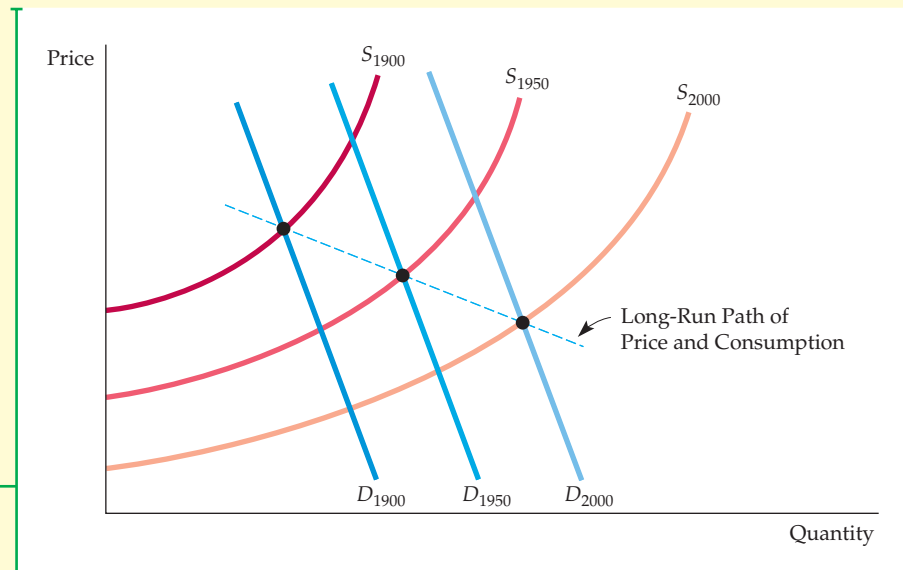
The answer is shown graphically in Figure 2.9. As you can see from that figure, the demands for these resources grew along with the world economy. But as demand grew, production costs fell. The decline in costs was due, first, to the discovery of new and bigger deposits that were cheaper to mine, and then to technical progress and the economic advantage of mining and refining on a large scale. As a result, the supply curve shifted over time to the right. Over the long term, because increases in supply were

⁴The index of U.S. copper consumption was around 102 in 1999 and 2000 but then dropped off significantly due to falling demand from 2001 to 2006. Consumption data (1880–1899) and price data (1880–1969) in Figure 2.8 are from Robert S. Manthey, *Natural Resource Commodities—A Century of Statistics* (Baltimore: Johns Hopkins University Press, 1978). More recent price (1970–2010) and consumption data (1970–2010) are from the U.S. Geological Survey—Minerals Information, Copper Statistics and Information (<http://minerals.usgs.gov/>).



FIGURE 2.9 LONG-RUN MOVEMENTS OF SUPPLY AND DEMAND FOR MINERAL RESOURCES

Although demand for most resources has increased dramatically over the past century, prices have fallen or risen only slightly in real (inflation-adjusted) terms because cost reductions have shifted the supply curve to the right just as dramatically.



greater than increases in demand, price often fell, as shown in Figure 2.9.

This is not to say that the prices of copper, iron, and coal will decline or remain constant forever. After all, these resources are *finite*. But as prices begin to rise, consumption will likely shift,

at least in part, to substitute materials. Copper, for example, has already been replaced in many applications by aluminum and, more recently, in electronic applications by fiber optics. (See Example 2.8 for a more detailed discussion of copper prices.)

EXAMPLE 2.4 THE EFFECTS OF 9/11 ON THE SUPPLY AND DEMAND FOR NEW YORK CITY OFFICE SPACE

The September 11, 2001, terrorist attack on the World Trade Center (WTC) complex damaged or destroyed 21 buildings, accounting for 31.2 million square feet (msf) of Manhattan office space—nearly 10 percent of the city's entire inventory. Just prior to the attack, the Manhattan office vacancy rate was 8.0 percent, and the average asking rent was \$52.50 per square foot (psf). Given the huge unexpected reduction in the quantity of office space supplied, we might expect the equilibrium rental price of office space to increase and, as a result, the equilibrium quantity of rented office space to decrease. And because it takes time to construct new office buildings and restore damaged ones, we might also expect the vacancy rate to decline sharply.

Surprisingly, however, the vacancy rate in Manhattan *increased* from 8.0 percent in August

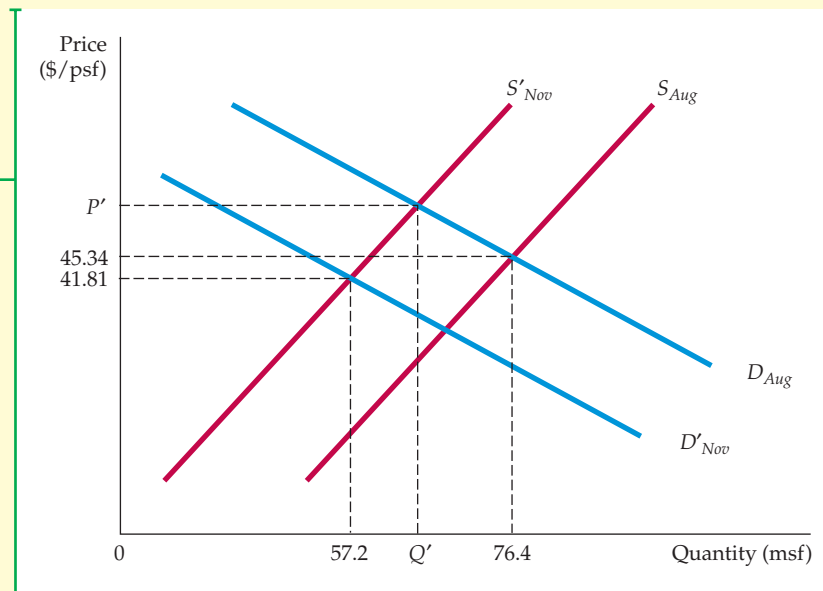
2001 to 9.3 percent in November 2001. Moreover, the average rental price *fell* from \$52.50 to \$50.75 per square foot. In downtown Manhattan, the location of the Trade Center, the changes were even more dramatic: The vacancy rate rose from 7.5 percent to 10.6 percent, and the average rental price fell nearly 8 percent, to \$41.81. What happened? Rental prices fell because the demand for office space fell.

Figure 2.10 describes the market for office space in downtown Manhattan. The supply and demand curves before 9/11 appear as S_{Aug} and D_{Aug} . The equilibrium price and quantity of downtown Manhattan office space were \$45.34 psf and 76.4 msf, respectively. The reduction in supply from August until November is indicated by a leftward shift in the supply curve (from S_{Aug} to S'_{Nov}); the result is a higher equilibrium price P' and a lower



FIGURE 2.10 SUPPLY AND DEMAND FOR NEW YORK CITY OFFICE SPACE

Following 9/11 the supply curve shifted to the left, but the demand curve also shifted to the left, so that the average rental price fell.



equilibrium quantity, Q' . This is the outcome that most forecasters predicted for the months following September 11.

Many forecasters, however, failed to predict the significant *decrease* in demand for office space complementing the loss in supply. First, many firms, both displaced and non-displaced, chose not to relocate downtown because of quality-of-life concerns (i.e., the WTC ruins, pollution, disabled transportation, and aging inventory). Firms displaced by the attack were also forced to reevaluate their office-space needs, and they ultimately repurchased a little more than 50 percent of their original office space in Manhattan. Others left Manhattan but stayed in New York City; still others moved to New Jersey.⁵ Furthermore, in late 2001, the U.S. economy was experiencing an economic slowdown (exacerbated by the events of September 11) that further reduced the demand for office space. Therefore, the cumulative decrease in demand (a shift from D_{Aug} to D'_{Nov}) actually caused the average rental price of downtown Manhattan office space to decrease rather than increase in the months following September 11. By November,

even though the price had fallen to \$41.81, there were 57.2 msf on the market.

There is evidence that office real estate markets in other major U.S. cities experienced similar surges in vacancy rates following 9/11. For instance, in Chicago, not only did vacancy rates increase in downtown office buildings, this increase was significantly more pronounced in properties in or near landmark buildings that are considered preferred targets for terrorist attacks.⁶

The Manhattan commercial real estate market bounced back strongly after 2001. In 2007, the office vacancy rate in Manhattan was 5.8 percent, its lowest figure since 9/11 and the average asking rent was over \$74 psf. By May 2009, the vacancy rate had risen above 13 percent. Financial services firms occupy more than a quarter of Manhattan office space, and with the financial crisis came a slump in commercial real estate. Goldman Sachs, for example, vacated more than 1 million square feet of office space. On the supply side, the new skyscraper at the northwest corner of the World Trade Center site will add 2.6 million square feet of office space upon completion.

⁵See Jason Bram, James Orr, and Carol Rapaport, "Measuring the Effects of the September 11 Attack on New York City," Federal Reserve Bank of New York, *Economic Policy Review*, November, 2002.

⁶See Alberto Abadie and Sofia Dermisi, "Is Terrorism Eroding Agglomeration Economies in Central Business Districts? Lessons from the Office Real Estate Market in Downtown Chicago," National Bureau of Economic Research, Working Paper 12678, November, 2006.



2.4 Elasticities of Supply and Demand

We have seen that the demand for a good depends not only on its price, but also on consumer income and on the prices of other goods. Likewise, supply depends both on price and on variables that affect production cost. For example, if the price of coffee increases, the quantity demanded will fall and the quantity supplied will rise. Often, however, we want to know *how much* the quantity supplied or demanded will rise or fall. How sensitive is the demand for coffee to its price? If price increases by 10 percent, how much will the quantity demanded change? How much will it change if income rises by 5 percent? We use *elasticities* to answer questions like these.

An **elasticity** measures the sensitivity of one variable to another. Specifically, it is a number that tells us *the percentage change that will occur in one variable in response to a 1-percent increase in another variable*. For example, the *price elasticity of demand* measures the sensitivity of quantity demanded to price changes. It tells us what the percentage change in the quantity demanded for a good will be following a 1-percent increase in the price of that good.

• **elasticity** Percentage change in one variable resulting from a 1-percent increase in another.

PRICE ELASTICITY OF DEMAND Let's look at this in more detail. We write the **price elasticity of demand**, E_p , as

$$E_p = (\% \Delta Q) / (\% \Delta P)$$

where $\% \Delta Q$ means "percentage change in quantity demanded" and $\% \Delta P$ means "percentage change in price." (The symbol Δ is the Greek capital letter *delta*; it means "the change in." So ΔX means "the change in the variable X ," say, from one year to the next.) The percentage change in a variable is just *the absolute change in the variable divided by the original level of the variable*. (If the Consumer Price Index were 200 at the beginning of the year and increased to 204 by the end of the year, the percentage change—or annual rate of inflation—would be $4/200 = .02$, or 2 percent.) Thus we can also write the price elasticity of demand as follows:⁷

• **price elasticity of demand** Percentage change in quantity demanded of a good resulting from a 1-percent increase in its price.

$$E_p = \frac{\Delta Q / Q}{\Delta P / P} = \frac{P \Delta Q}{Q \Delta P} \quad (2.1)$$

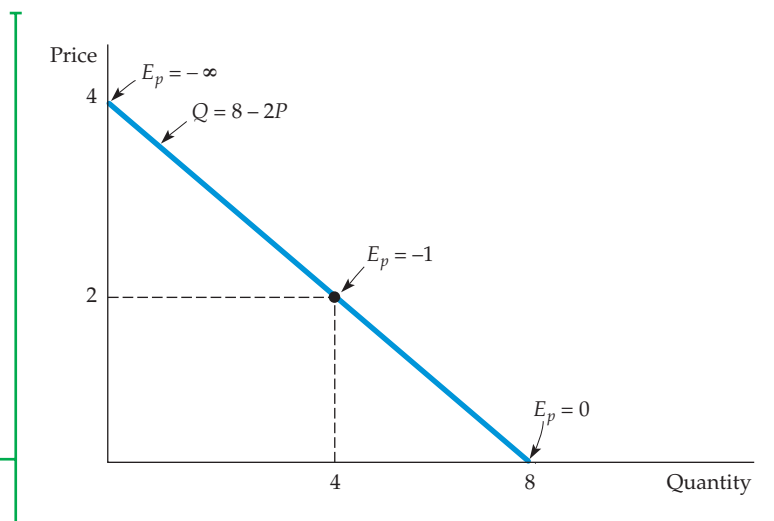
The price elasticity of demand is usually a negative number. When the price of a good increases, the quantity demanded usually falls. Thus $\Delta Q / \Delta P$ (the change in quantity for a change in price) is negative, as is E_p . Sometimes we refer to the *magnitude* of the price elasticity—i.e., its absolute size. For example, if $E_p = -2$, we say that the elasticity is 2 in magnitude.

When the price elasticity is greater than 1 in magnitude, we say that demand is *price elastic* because the percentage decline in quantity demanded is greater than the percentage increase in price. If the price elasticity is less than 1 in magnitude, demand is said to be *price inelastic*. In general, the price elasticity of demand for a good depends on the availability of other goods that can be substituted for it. When there are close substitutes, a price increase will cause the consumer to buy less of the good and more of the substitute. Demand will then be highly price elastic. When there are no close substitutes, demand will tend to be price inelastic.

⁷In terms of infinitesimal changes (letting the ΔP become very small), $E_p = (P/Q)(dQ/dP)$.

FIGURE 2.11 LINEAR DEMAND CURVE

The price elasticity of demand depends not only on the slope of the demand curve but also on the price and quantity. The elasticity, therefore, varies along the curve as price and quantity change. Slope is constant for this linear demand curve. Near the top, because price is high and quantity is small, the elasticity is large in magnitude. The elasticity becomes smaller as we move down the curve.



LINEAR DEMAND CURVE Equation (2.1) says that the price elasticity of demand is the change in quantity associated with a change in price ($\Delta Q / \Delta P$) times the ratio of price to quantity (P / Q). But as we move down the demand curve, $\Delta Q / \Delta P$ may change, and the price and quantity will always change. Therefore, the price elasticity of demand must be measured *at a particular point on the demand curve* and will generally change as we move along the curve.

This principle is easiest to see for a **linear demand curve**—that is, a demand curve of the form

$$Q = a - bP$$

As an example, consider the demand curve

$$Q = 8 - 2P$$

For this curve, $\Delta Q / \Delta P$ is constant and equal to -2 (a ΔP of 1 results in a ΔQ of -2). However, the curve does *not* have a constant elasticity. Observe from Figure 2.11 that as we move down the curve, the ratio P / Q falls; the elasticity therefore decreases in magnitude. Near the intersection of the curve with the price axis, Q is very small, so $E_p = -2(P / Q)$ is large in magnitude. When $P = 2$ and $Q = 4$, $E_p = -1$. At the intersection with the quantity axis, $P = 0$ so $E_p = 0$.

Because we draw demand (and supply) curves with price on the vertical axis and quantity on the horizontal axis, $\Delta Q / \Delta P = (1 / \text{slope of curve})$. As a result, for any price and quantity combination, the steeper the slope of the curve, the less elastic is demand. Figure 2.12 shows two special cases. Figure 2.12(a) shows a demand curve reflecting **infinitely elastic demand**: Consumers will buy as much as they can at a single price P^* . For even the smallest increase in price above this level, quantity demanded drops to zero, and for any decrease in price, quantity demanded increases without limit. The demand curve in Figure 2.12(b), on the other hand, reflects **completely inelastic demand**: Consumers will buy a fixed quantity Q^* , no matter what the price.

OTHER DEMAND ELASTICITIES We will also be interested in elasticities of demand with respect to other variables besides price. For example, demand for most goods usually rises when aggregate income rises. The **income elasticity of**

- **linear demand curve**

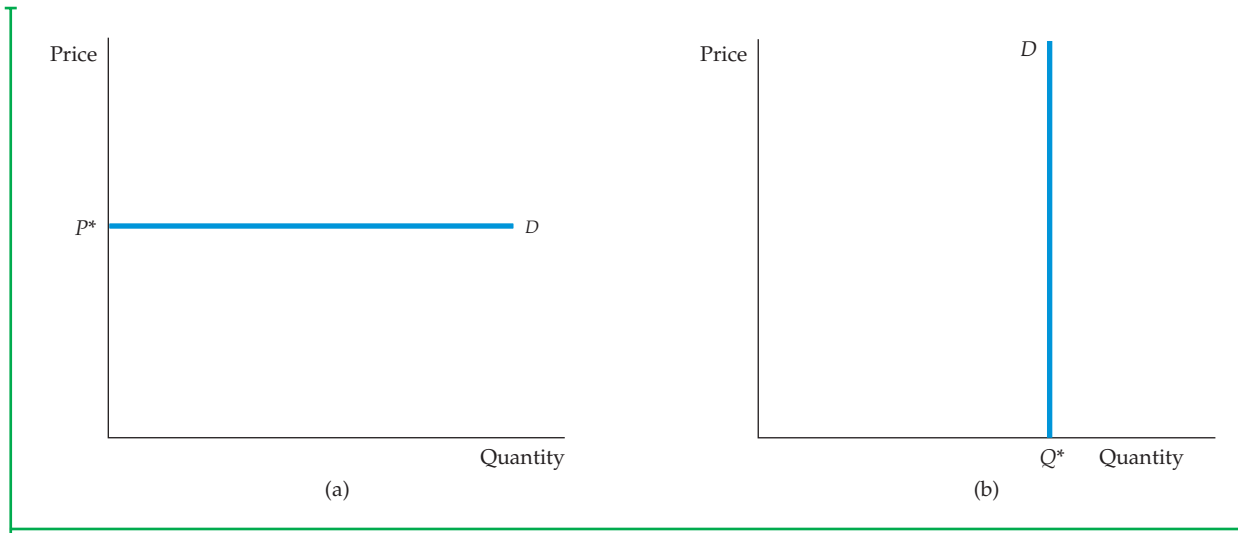
Demand curve that is a straight line.

- **infinitely elastic demand**

Principle that consumers will buy as much of a good as they can get at a single price, but for any higher price the quantity demanded drops to zero, while for any lower price the quantity demanded increases without limit.

- **completely inelastic demand**

Principle that consumers will buy a fixed quantity of a good regardless of its price.

**FIGURE 2.12****(a) INFINITELY ELASTIC DEMAND (b) COMPLETELY INELASTIC DEMAND**

(a) For a horizontal demand curve, $\Delta Q/\Delta P$ is infinite. Because a tiny change in price leads to an enormous change in demand, the elasticity of demand is infinite. (b) For a vertical demand curve, $\Delta Q/\Delta P$ is zero. Because the quantity demanded is the same no matter what the price, the elasticity of demand is zero.

demand is the percentage change in the quantity demanded, Q , resulting from a 1-percent increase in income I :

$$E_I = \frac{\Delta Q/Q}{\Delta I/I} = \frac{I}{Q} \frac{\Delta Q}{\Delta I} \quad (2.2)$$

• **income elasticity of demand** Percentage change in the quantity demanded resulting from a 1-percent increase in income.

The demand for some goods is also affected by the prices of other goods. For example, because butter and margarine can easily be substituted for each other, the demand for each depends on the price of the other. A **cross-price elasticity of demand** refers to the percentage change in the quantity demanded for a good that results from a 1-percent increase in the price of another good. So the elasticity of demand for butter with respect to the price of margarine would be written as

$$E_{Q_b P_m} = \frac{\Delta Q_b/Q_b}{\Delta P_m/P_m} = \frac{P_m}{Q_b} \frac{\Delta Q_b}{\Delta P_m} \quad (2.3)$$

• **cross-price elasticity of demand** Percentage change in the quantity demanded of one good resulting from a 1-percent increase in the price of another.

where Q_b is the quantity of butter and P_m is the price of margarine.

In this example, the cross-price elasticities will be positive because the goods are *substitutes*: Because they compete in the market, a rise in the price of margarine, which makes butter cheaper relative to margarine, leads to an increase in the quantity of butter demanded. (Because the demand curve for butter will shift to the right, the price of butter will rise.) But this is not always the case. Some goods are *complements*: Because they tend to be used together, an increase in the price of one tends to push down the consumption of the other. Take gasoline and motor oil. If the price of gasoline goes up, the quantity of



• **price elasticity of supply**
Percentage change in quantity supplied resulting from a 1-percent increase in price.

gasoline demanded falls—motorists will drive less. And because people are driving less, the demand for motor oil also falls. (The entire demand curve for motor oil shifts to the left.) Thus, the cross-price elasticity of motor oil with respect to gasoline is negative.

ELASTICITIES OF SUPPLY Elasticities of supply are defined in a similar manner. The **price elasticity of supply** is the percentage change in the quantity supplied resulting from a 1-percent increase in price. This elasticity is usually positive because a higher price gives producers an incentive to increase output.

We can also refer to elasticities of supply with respect to such variables as interest rates, wage rates, and the prices of raw materials and other intermediate goods used to manufacture the product in question. For example, for most manufactured goods, the elasticities of supply with respect to the prices of raw materials are negative. An increase in the price of a raw material input means higher costs for the firm; other things being equal, therefore, the quantity supplied will fall.

Point versus Arc Elasticities

• **point elasticity of demand**
Price elasticity at a particular point on the demand curve.

So far, we have considered elasticities at a particular point on the demand curve or the supply curve. These are called *point elasticities*. The **point elasticity of demand**, for example, is *the price elasticity of demand at a particular point on the demand curve* and is defined by Equation (2.1). As we demonstrated in Figure 2.11 using a linear demand curve, the point elasticity of demand can vary depending on where it is measured along the demand curve.

There are times, however, when we want to calculate a price elasticity over some portion of the demand curve (or supply curve) rather than at a single point. Suppose, for example, that we are contemplating an increase in the price of a product from \$8.00 to \$10.00 and expect the quantity demanded to fall from 6 units to 4. How should we calculate the price elasticity of demand? Is the price increase 25 percent (a \$2 increase divided by the original price of \$8), or is it 20 percent (a \$2 increase divided by the new price of \$10)? Is the percentage decrease in quantity demanded 33 1/3 percent (2/6) or 50 percent (2/4)?

There is no correct answer to such questions. We could calculate the price elasticity using the original price and quantity. If so, we would find that $E_p = (-33\frac{1}{3} \text{ percent} / 25 \text{ percent}) = -1.33$. Or we could use the new price and quantity, in which case we would find that $E_p = (-50 \text{ percent} / 20 \text{ percent}) = -2.5$. The difference between these two calculated elasticities is large, and neither seems preferable to the other.

• **arc elasticity of demand**
Price elasticity calculated over a range of prices.

ARC ELASTICITY OF DEMAND We can resolve this problem by using the **arc elasticity of demand**: *the elasticity calculated over a range of prices*. Rather than choose either the initial or the final price, we use an average of the two, \bar{P} ; for the quantity demanded, we use \bar{Q} . Thus the arc elasticity of demand is given by

$$\text{Arc elasticity: } E_p = (\Delta Q / \Delta P)(\bar{P} / \bar{Q}) \quad (2.4)$$

In our example, the average price is \$9 and the average quantity 5 units. Thus the arc elasticity is

$$E_p = (-2 / \$2)(\$9 / 5) = -1.8$$



The arc elasticity will always lie somewhere (but not necessarily halfway) between the point elasticities calculated at the lower and the higher prices.

Although the arc elasticity of demand is sometimes useful, economists generally use the word “elasticity” to refer to a *point* elasticity. Throughout the rest of this book, we will do the same, unless noted otherwise.

EXAMPLE 2.5 THE MARKET FOR WHEAT

Wheat is an important agricultural commodity, and the wheat market has been studied extensively by agricultural economists. During recent decades, changes in the wheat market had major implications for both American farmers and U.S. agricultural policy. To understand what happened, let’s examine the behavior of supply and demand beginning in 1981.



From statistical studies, we know that for 1981 the supply curve for wheat was approximately as follows:⁸

$$\text{Supply: } Q_S = 1800 + 240P$$

where price is measured in nominal dollars per bushel and quantities in millions of bushels per year. These studies also indicate that in 1981, the demand curve for wheat was

$$\text{Demand: } Q_D = 3550 - 266P$$

By setting the quantity supplied equal to the quantity demanded, we can determine the market-clearing price of wheat for 1981:

$$\begin{aligned} Q_S &= Q_D \\ 1800 + 240P &= 3550 - 266P \\ 506P &= 1750 \\ P &= \$3.46 \text{ per bushel} \end{aligned}$$

To find the market-clearing quantity, substitute this price of \$3.46 into either the supply curve equation or the demand curve equation. Substituting into the supply curve equation, we get

$$Q = 1800 + (240)(3.46) = 2630 \text{ million bushels}$$

⁸For a survey of statistical studies of the demand and supply of wheat and an analysis of evolving market conditions, see Larry Salathe and Sudchada Langley, “An Empirical Analysis of Alternative Export Subsidy Programs for U.S. Wheat,” *Agricultural Economics Research* 38:1 (Winter 1986). The supply and demand curves in this example are based on the studies they surveyed.



What are the price elasticities of demand and supply at this price and quantity? We use the demand curve to find the price elasticity of demand:

$$E_P^D = \frac{P}{Q} \frac{\Delta Q_D}{\Delta P} = \frac{3.46}{2630} (-266) = -0.35$$

Thus demand is inelastic. We can likewise calculate the price elasticity of supply:

$$\begin{aligned} E_P^S &= \frac{P}{Q} \frac{\Delta Q_S}{\Delta P} \\ &= \frac{3.46}{2630} (240) = 0.32 \end{aligned}$$

Because these supply and demand curves are linear, the price elasticities will vary as we move along the curves. For example, suppose that a drought caused the supply curve to shift far enough to the left to push the price up to \$4.00 per bushel. In this case, the quantity demanded would fall to $3550 - (266)(4.00) = 2486$ million bushels. At this price and quantity, the elasticity of demand would be

$$E_P^D = \frac{4.00}{2486} (-266) = -0.43$$

The wheat market has evolved over the years, in part because of changes in demand. The demand for wheat has two components: domestic (demand by U.S. consumers) and export (demand by foreign consumers). During the 1980s and 1990s, domestic demand for wheat rose only slightly (due to modest increases in population and income). Export demand, however, fell sharply. There were several reasons. First and foremost was the success of the Green Revolution in agriculture: Developing countries like India, which had been large importers of wheat, became increasingly self-sufficient. In addition, European countries adopted protectionist policies that subsidized their own production and imposed tariff barriers against imported wheat.

In 2007, demand and supply were

$$\text{Demand: } Q_D = 2900 - 125P$$

$$\text{Supply: } Q_S = 1460 + 115P$$

Once again, equating quantity supplied and quantity demanded yields the market-clearing (nominal) price and quantity:

$$1460 + 115P = 2900 - 125P$$

$$P = \$6.00 \text{ per bushel}$$

$$Q = 1460 + (115)(6) = 2150 \text{ million bushels}$$



Thus the price of wheat (in nominal terms) rose considerably since 1981. In fact, nearly all of this increase occurred during 2005 to 2007. (In 2002, for example, the price of wheat was only \$2.78 per bushel.) The causes? Dry weather in 2005, even dryer weather in 2006, and heavy rains in 2007 combined with increased export demand. You can check to see that, at the 2007 price and quantity, the price elasticity of demand was -0.35 and the price elasticity of supply 0.32 . Given these low elasticities, it is not surprising that the price of wheat rose so sharply.⁹

International demand for U. S. wheat fluctuates with the weather and political conditions in other major wheat producing countries, such as China, India and Russia. Between 2008 and 2010, U.S. wheat exports fell by 30% in the face of robust international production, so the price of wheat reached a low of \$4.87 in 2010, down from \$6.48 two years earlier. Inclement weather led to shortfalls in 2011, however, and U.S. exports shot up by 33%, driving the price up to \$5.70 in 2011.

We found that the market-clearing price of wheat was \$3.46 in 1981, but in fact the price was greater than this. Why? Because the U.S. government bought wheat through its price support program. In addition, farmers have been receiving direct subsidies for the wheat they produce. This aid to farmers (at the expense of taxpayers) has increased in magnitude. In 2002—and again in 2008—Congress passed legislation continuing (and in some cases expanding) subsidies to farmers. The Food, Conservation, and Energy Act of 2008 authorized farm aid through 2012, at a projected cost of \$284 billion over five years. Recent U.S. budget crises, however, have given support to those in Congress who feel these subsidies should end.¹⁰

Agricultural policies that support farmers exist in the United States, Europe, Japan, and many other countries. We discuss how these policies work, and evaluate the costs and benefits for consumers, farmers, and the government budget in Chapter 9.

2.5 Short-Run versus Long-Run Elasticities

When analyzing demand and supply, we must distinguish between the short run and the long run. In other words, if we ask how much demand or supply changes in response to a change in price, we must be clear about *how much time is allowed to pass before we measure the changes in the quantity demanded or supplied*. If we allow only a short time to pass—say, one year or less—then we are dealing with the *short run*. When we refer to the *long run* we mean that enough time is allowed for consumers or producers to *adjust fully* to the price change. In general, short-run demand and supply curves look very different from their long-run counterparts.

⁹These are short-run elasticity estimates from Economics Research Service (ERS) of the U.S. Department of Agriculture (USDA). For more information, consult the following publications: William Lin, Paul C. Westcott, Robert Skinner, Scott Sanford, and Daniel G. De La Torre Ugarte, *Supply Response Under the 1996 Farm Act and Implications for the U.S. Field Crops Sector* (Technical Bulletin No. 1888, ERS, USDA, July 2000, <http://www.ers.usda.gov/>); and James Barnes and Dennis Shields, *The Growth in U.S. Wheat Food Demand* (Wheat Situation and Outlook Yearbook, WHS-1998, <http://www.ers.usda.gov/>).

¹⁰For more information on past farm bills: <http://www.ers.usda.gov/farmbill/2008/>.



Demand

For many goods, demand is much more price elastic in the long run than in the short run. For one thing, it takes time for people to change their consumption habits. For example, even if the price of coffee rises sharply, the quantity demanded will fall only gradually as consumers begin to drink less. In addition, the demand for a good might be linked to the stock of another good that changes only slowly. For example, the demand for gasoline is much more elastic in the long run than in the short run. A sharply higher price of gasoline reduces the quantity demanded in the short run by causing motorists to drive less, but it has its greatest impact on demand by inducing consumers to buy smaller and more fuel-efficient cars. But because the stock of cars changes only slowly, the quantity of gasoline demanded falls only slowly. Figure 2.13(a) shows short-run and long-run demand curves for goods such as these.

DEMAND AND DURABILITY On the other hand, for some goods just the opposite is true—demand is more elastic in the short run than in the long run. Because these goods (automobiles, refrigerators, televisions, or the capital equipment purchased by industry) are *durable*, the total stock of each good owned by

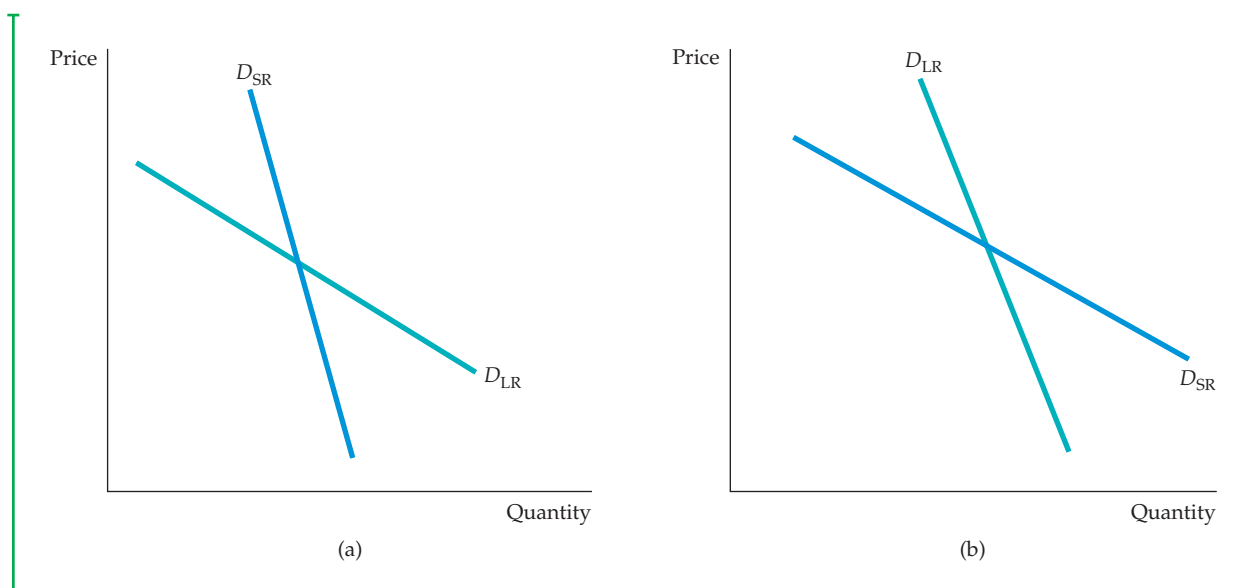


FIGURE 2.13

(a) GASOLINE: SHORT-RUN AND LONG-RUN DEMAND CURVES

(b) AUTOMOBILES: SHORT-RUN AND LONG-RUN DEMAND CURVES

(a) In the short run, an increase in price has only a small effect on the quantity of gasoline demanded. Motorists may drive less, but they will not change the kinds of cars they are driving overnight. In the longer run, however, because they will shift to smaller and more fuel-efficient cars, the effect of the price increase will be larger. Demand, therefore, is more elastic in the long run than in the short run.

(b) The opposite is true for automobile demand. If price increases, consumers initially defer buying new cars; thus annual quantity demanded falls sharply. In the longer run, however, old cars wear out and must be replaced; thus annual quantity demanded picks up. Demand, therefore, is less elastic in the long run than in the short run.



consumers is large relative to annual production. As a result, a small change in the total stock that consumers want to hold can result in a large percentage change in the level of purchases.

Suppose, for example, that the price of refrigerators goes up 10 percent, causing the total stock of refrigerators that consumers want to hold to drop 5 percent. Initially, this will cause purchases of new refrigerators to drop much more than 5 percent. But eventually, as consumers' refrigerators depreciate (and units must be replaced), the quantity demanded will increase again. In the long run, the total stock of refrigerators owned by consumers will be about 5 percent less than before the price increase. In this case, while the long-run price elasticity of demand for refrigerators would be $-.05/.10 = -0.5$, the short-run elasticity would be much larger in magnitude.

Or consider automobiles. Although annual U.S. demand—new car purchases—is about 10 to 12 million, the stock of cars that people own is around 130 million. If automobile prices rise, many people will delay buying new cars. The quantity demanded will fall sharply, even though the total stock of cars that consumers might want to own at these higher prices falls only a small amount. Eventually, however, because old cars wear out and must be replaced, the quantity of new cars demanded picks up again. As a result, the long-run change in the quantity demanded is much smaller than the short-run change. Figure 2.13(b) shows demand curves for a durable good like automobiles.

INCOME ELASTICITIES Income elasticities also differ from the short run to the long run. For most goods and services—foods, beverages, fuel, entertainment, etc.—the income elasticity of demand is larger in the long run than in the short run. Consider the behavior of gasoline consumption during a period of strong economic growth during which aggregate income rises by 10 percent. Eventually people will increase gasoline consumption because they can afford to take more trips and perhaps own larger cars. But this change in consumption takes time, and demand initially increases only by a small amount. Thus, the long-run elasticity will be larger than the short-run elasticity.

For a durable good, the opposite is true. Again, consider automobiles. If aggregate income rises by 10 percent, the total stock of cars that consumers will want to own will also rise—say, by 5 percent. But this change means a much larger increase in *current purchases* of cars. (If the stock is 130 million, a 5-percent increase is 6.5 million, which might be about 60 to 70 percent of normal demand in a single year.) Eventually consumers succeed in increasing the total number of cars owned; after the stock has been rebuilt, new purchases are made largely to replace old cars. (These new purchases will still be greater than before because a larger stock of cars outstanding means that more cars need to be replaced each year.) Clearly, the short-run income elasticity of demand will be much larger than the long-run elasticity.

CYCLICAL INDUSTRIES Because the demands for durable goods fluctuate so sharply in response to short-run changes in income, the industries that produce these goods are quite vulnerable to changing macroeconomic conditions, and in particular to the business cycle—recessions and booms. Thus, these industries are often called **cyclical industries**—their sales patterns tend

• **cyclical industries** Industries in which sales tend to magnify cyclical changes in gross domestic product and national income.

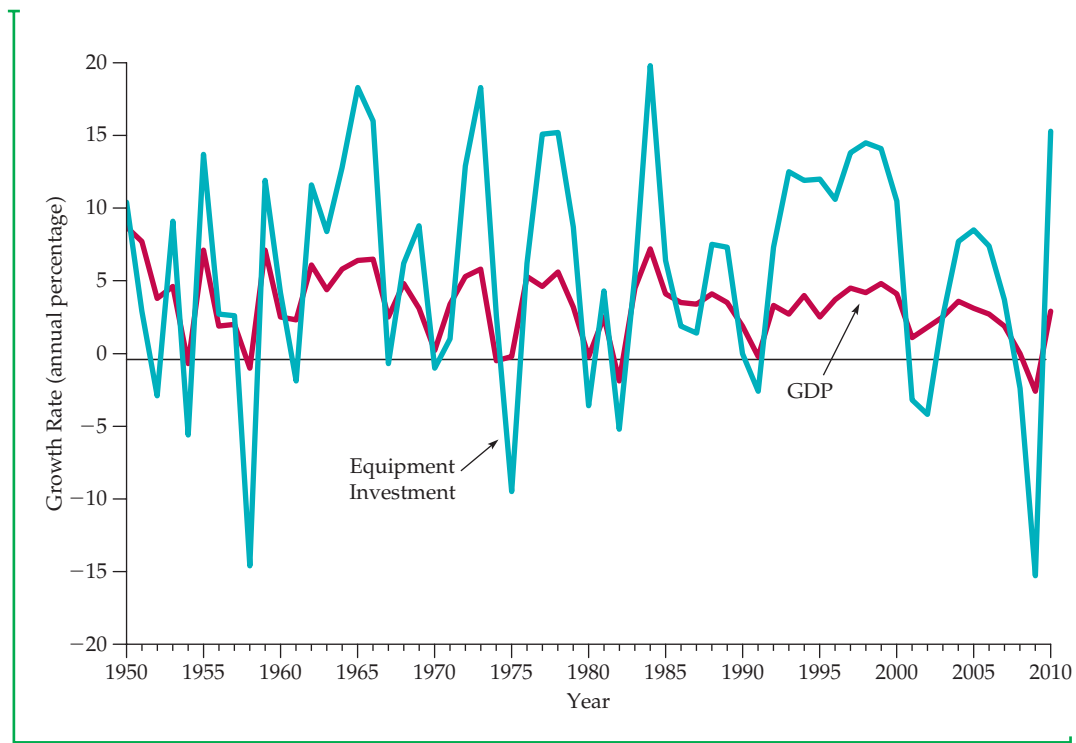


FIGURE 2.14
GDP AND INVESTMENT IN DURABLE EQUIPMENT

Annual growth rates are compared for GDP and investment in durable equipment. Because the short-run GDP elasticity of demand is larger than the long-run elasticity for long-lived capital equipment, changes in investment in equipment magnify changes in GDP. Thus capital goods industries are considered “cyclical.”

to magnify cyclical changes in gross domestic product (GDP) and national income.

Figures 2.14 and 2.15 illustrate this principle. Figure 2.14 plots two variables over time: the annual real (inflation-adjusted) rate of growth of GDP and the annual real rate of growth of investment in producers’ durable equipment (i.e., machinery and other equipment purchased by firms). Note that although the durable equipment series follows the same pattern as the GDP series, the changes in GDP are magnified. For example, in 1961–1966 GDP grew by at least 4 percent each year. Purchases of durable equipment also grew, but by much more (over 10 percent in 1963–1966). Equipment investment likewise grew much more quickly than GDP during 1993–1998. On the other hand, during the recessions of 1974–1975, 1982, 1991, 2001, and 2008, equipment purchases fell by much more than GDP.

Figure 2.15 also shows the real rate of growth of GDP, along with the annual real rates of growth of spending by consumers on durable goods (automobiles, appliances, etc.) and nondurable goods (food, fuel, clothing, etc.). Note that while both consumption series follow GDP, only the durable goods series tends to magnify changes in GDP. Changes in consumption of nondurables are roughly the same as changes in GDP, but changes in consumption of durables are usually several times larger. This is why companies

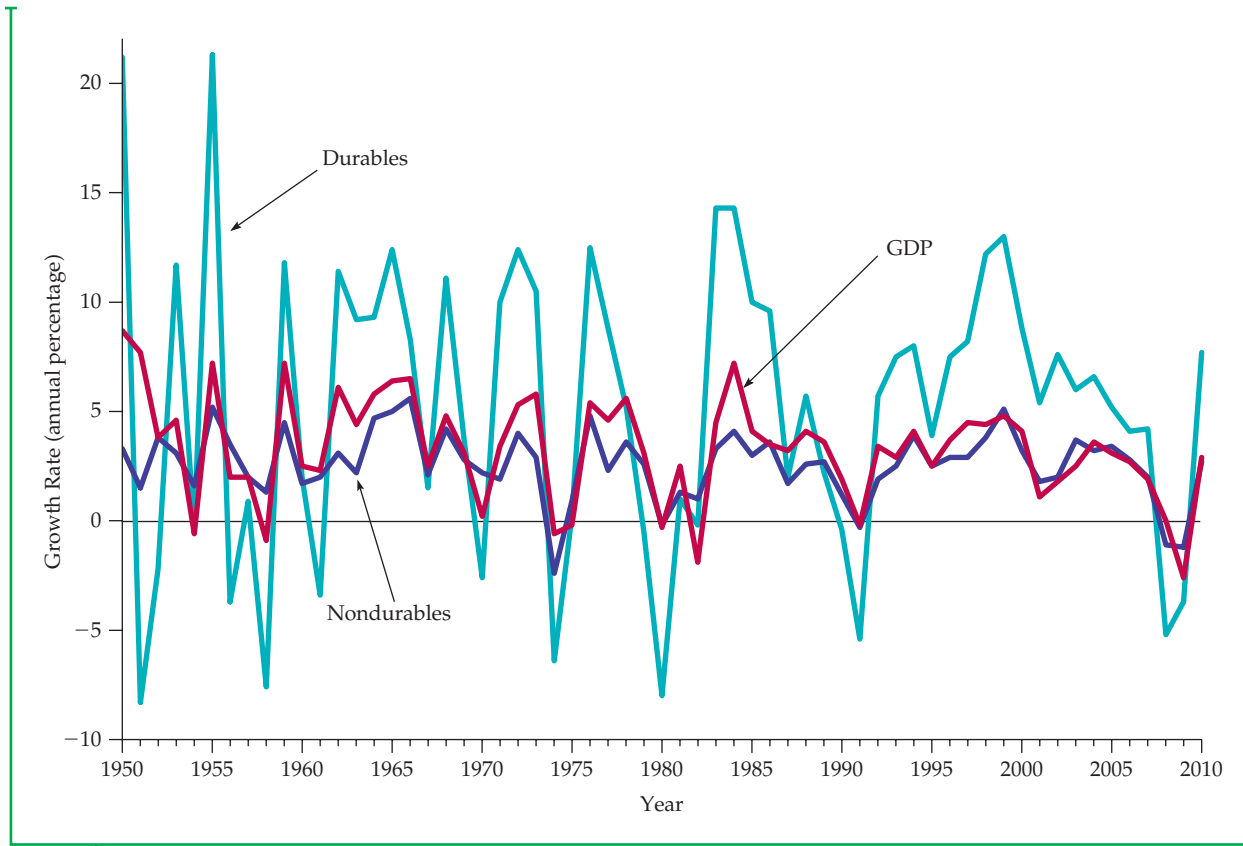


FIGURE 2.15
CONSUMPTION OF DURABLES VERSUS NONDURABLES

Annual growth rates are compared for GDP, consumer expenditures on durable goods (automobiles, appliances, furniture, etc.), and consumer expenditures on nondurable goods (food, clothing, services, etc.). Because the stock of durables is large compared with annual demand, short-run demand elasticities are larger than long-run elasticities. Like capital equipment, industries that produce consumer durables are “cyclical” (i.e., changes in GDP are magnified). This is not true for producers of nondurables.

such as General Motors and General Electric are considered “cyclical”: Sales of cars and electrical appliances are strongly affected by changing macroeconomic conditions.

EXAMPLE 2.6 THE DEMAND FOR GASOLINE AND AUTOMOBILES

Gasoline and automobiles exemplify some of the different characteristics of demand discussed above. They are complementary goods—an increase in the price of one tends to reduce the demand for the other. In addition, their respective dynamic behaviors (long-run versus short-run elasticities) are just the opposite from each other. For gasoline, the long-run price and income elasticities are larger than the short-run elasticities; for automobiles, the reverse is true.

**TABLE 2.1 DEMAND FOR GASOLINE**

ELASTICITY	NUMBER OF YEARS ALLOWED TO PASS FOLLOWING A PRICE OR INCOME CHANGE				
	1	2	3	5	10
Price	-0.2	-0.3	-0.4	-0.5	-0.8
Income	0.2	0.4	0.5	0.6	1.0

There have been a number of statistical studies of the demands for gasoline and automobiles. Here we report elasticity estimates based on several that emphasize the dynamic response of demand.¹¹ Table 2.1 shows price and income elasticities of demand for gasoline in the United States for the short run, the long run, and just about everything in between.

Note the large differences between the long-run and the short-run elasticities. Following the sharp increases that occurred in the price of gasoline with the rise of the OPEC oil cartel in 1974, many people (including executives in the automobile and oil industries) claimed that the quantity of gasoline demanded would not change much—that demand was not very elastic. Indeed, for the first year after the price rise, they were right. But demand did eventually change. It just took time for people to alter their driving habits and to replace large cars with smaller and more fuel-efficient ones. This response continued after the second sharp increase in oil prices that occurred in 1979–1980. It was partly because of this response that OPEC could not maintain oil prices above \$30 per barrel, and prices fell. The oil and gasoline price increases that occurred in 2005–2011 likewise led to a gradual demand response.

Table 2.2 shows price and income elasticities of demand for automobiles. Note that the short-run elasticities are much larger than the long-run

TABLE 2.2 DEMAND FOR AUTOMOBILES

ELASTICITY	NUMBER OF YEARS ALLOWED TO PASS FOLLOWING A PRICE OR INCOME CHANGE				
	1	2	3	5	10
Price	-1.2	-0.9	-0.8	-0.6	-0.4
Income	3.0	2.3	1.9	1.4	1.0

¹¹For gasoline and automobile demand studies and elasticity estimates, see R. S. Pindyck, *The Structure of World Energy Demand* (Cambridge, MA: MIT Press, 1979); Carol Dahl and Thomas Sterner, "Analyzing Gasoline Demand Elasticities: A Survey," *Energy Economics* (July 1991); Molly Espey, "Gasoline Demand Revised: An International Meta-Analysis of Elasticities," *Energy Economics* (July 1998); David L. Greene, James R. Kahn, and Robert C. Gibson, "Fuel Economy Rebound Effects for U.S. Household Vehicles," *The Energy Journal* 20 (1999); Daniel Graham and Stephen Glaister, "The Demand for Automobile Fuel: A Survey of Elasticities," *Journal of Transport Economics and Policy* 36 (January 2002); and Ian Parry and Kenneth Small, "Does Britain or the United States Have the Right Gasoline Tax?" *American Economic Review* 95 (2005).



elasticities. It should be clear from the income elasticities why the automobile industry is so highly cyclical. For example, GDP fell 2 percent in real (inflation-adjusted) terms during the 1991 recession, but automobile sales fell by about 8 percent. Auto sales began to recover in 1993, and rose sharply between 1995 and 1999. During the 2008 recession, GDP fell by nearly 3 percent, and car and truck sales decreased by 21%. Sales began to recover in 2010, when they increased by nearly 10%.

Supply

Elasticities of supply also differ from the long run to the short run. For most products, long-run supply is much more price elastic than short-run supply: Firms face *capacity constraints* in the short run and need time to expand capacity by building new production facilities and hiring workers to staff them. This is not to say that the quantity supplied will not increase in the short run if price goes up sharply. Even in the short run, firms can increase output by using their existing facilities for more hours per week, paying workers to work overtime, and hiring some new workers immediately. But firms will be able to expand output much more when they have the time to expand their facilities and hire larger permanent workforces.

For some goods and services, short-run supply is completely inelastic. Rental housing in most cities is an example. In the very short run, there is only a fixed number of rental units. Thus an increase in demand only pushes rents up. In the longer run, and without rent controls, higher rents provide an incentive to renovate existing buildings and construct new ones. As a result, the quantity supplied increases.

For most goods, however, firms can find ways to increase output even in the short run—if the price incentive is strong enough. However, because various constraints make it costly to increase output rapidly, it may require large price increases to elicit small short-run increases in the quantity supplied. We discuss these characteristics of supply in more detail in Chapter 8.

SUPPLY AND DURABILITY For some goods, supply is more elastic in the short run than in the long run. Such goods are durable and can be recycled as part of supply if price goes up. An example is the *secondary supply* of metals: the supply from *scrap metal*, which is often melted down and refabricated. When the price of copper goes up, it increases the incentive to convert scrap copper into new supply, so that, initially, secondary supply increases sharply. Eventually, however, the stock of good-quality scrap falls, making the melting, purifying, and refabricating more costly. Secondary supply then contracts. Thus the long-run price elasticity of secondary supply is smaller than the short-run elasticity.

Figures 2.16(a) and 2.16(b) show short-run and long-run supply curves for primary (production from the mining and smelting of ore) and secondary copper production. Table 2.3 shows estimates of the elasticities for each component of supply and for total supply, based on a weighted average of the component elasticities.¹² Because secondary supply is only about 20 percent of total supply, the price elasticity of total supply is larger in the long run than in the short run.

¹²These estimates were obtained by aggregating the regional estimates reported in Franklin M. Fisher, Paul H. Cootner, and Martin N. Baily, “An Econometric Model of the World Copper Industry,” *Bell Journal of Economics* 3 (Autumn 1972): 568–609.



FIGURE 2.16
COPPER: SHORT-RUN AND LONG-RUN SUPPLY CURVES

Like that of most goods, the supply of primary copper, shown in part (a), is more elastic in the long run. If price increases, firms would like to produce more but are limited by capacity constraints in the short run. In the longer run, they can add to capacity and produce more. Part (b) shows supply curves for secondary copper. If the price increases, there is a greater incentive to convert scrap copper into new supply. Initially, therefore, secondary supply (i.e., supply from scrap) increases sharply. But later, as the stock of scrap falls, secondary supply contracts. Secondary supply is therefore less elastic in the long run than in the short run.

TABLE 2.3 SUPPLY OF COPPER

PRICE ELASTICITY OF:	SHORT-RUN	LONG-RUN
Primary supply	0.20	1.60
Secondary supply	0.43	0.31
Total supply	0.25	1.50

EXAMPLE 2.7 THE WEATHER IN BRAZIL AND THE PRICE OF COFFEE IN NEW YORK

Droughts or subfreezing weather occasionally destroy or damage many of Brazil's coffee trees. Because Brazil is by far the world's largest coffee producer the result is a decrease in the supply of coffee and a sharp run-up in its price.

In July 1975, for example, a frost destroyed most of Brazil's 1976–1977 coffee crop. (Remember that it is winter



Brazil's crop. Finally, starting in June 1994, freezing

in Brazil when it is summer in the northern hemisphere.) As Figure 2.17 shows, the price of a pound of coffee in New York went from 68 cents in 1975 to \$1.23 in 1976 and \$2.70 in 1977. Prices fell but then jumped again in 1986, after a seven-month drought in 1985 ruined much of

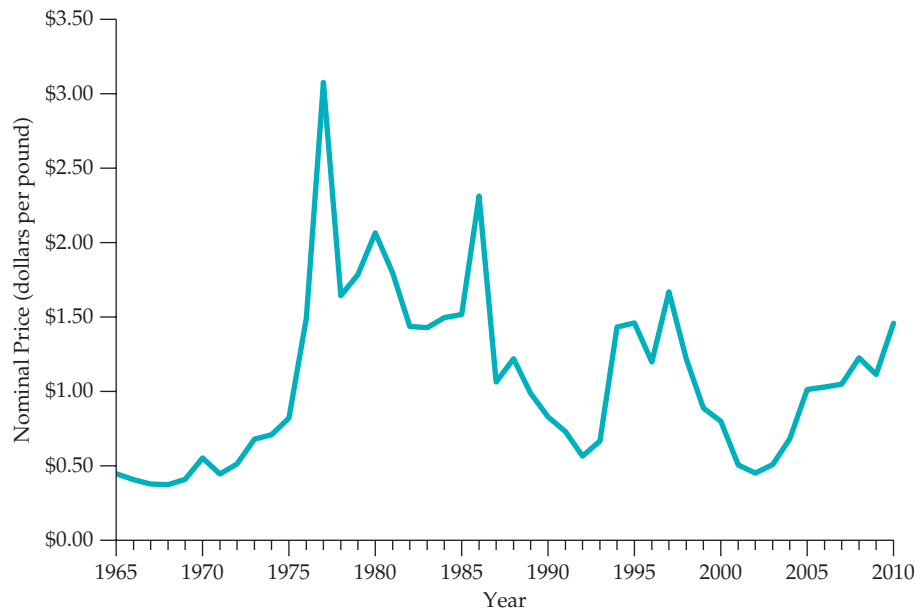


FIGURE 2.17
PRICE OF BRAZILIAN COFFEE

When droughts or freezes damage Brazil's coffee trees, the price of coffee can soar. The price usually falls again after a few years, as demand and supply adjust.

weather followed by a drought destroyed nearly half of Brazil's crop. As a result, the price of coffee in 1994–1995 was about double its 1993 level. By 2002, however, the price had dropped to its lowest level in 30 years. (Researchers predict that over the next 50 years, global warming may eliminate as much as 60 percent of Brazil's coffee-growing areas, resulting in a major decline in coffee production and an increase in prices. Should that happen, we will discuss it in the twentieth edition of this book.)

The important point in Figure 2.17 is that any run-up in price following a freeze or drought is usually short-lived. Within a year, price begins to fall; within three or four years, it returns to its earlier levels. In 1978, for example, the price of coffee in New York fell to \$1.48 per pound, and by 1983, it had fallen in real (inflation-adjusted) terms to within a few cents of its prefreeze 1975 price.¹³ Likewise, in 1987 the price of coffee fell to below its predrought

1984 level, and then continued declining until the 1994 freeze. After hitting a low of 45 cents per pound in 2002, coffee prices increased at an average rate of 17% per year, reaching \$1.46—equal to the 1995 peak—in 2010. Brazilian coffee growers have worked to increase their production in the past decade, but bad weather has led to inconsistent crop yields.

Coffee prices behave this way because both demand and supply (especially supply) are much more elastic in the long run than in the short run. Figure 2.18 illustrates this fact. Note from part (a) of the figure that in the very short run (within one or two months after a freeze), supply is completely inelastic: There are simply a fixed number of coffee beans, some of which have been damaged by the frost. Demand is also relatively inelastic. As a result of the frost, the supply curve shifts to the left, and price increases sharply, from P_0 to P_1 .

¹³During 1980, however, prices temporarily went just above \$2.00 per pound as a result of export quotas imposed under the International Coffee Agreement (ICA). The ICA is essentially a cartel agreement implemented by the coffee-producing countries in 1968. It has been largely ineffective and has seldom had an effect on the price. We discuss cartel pricing in detail in Chapter 12.

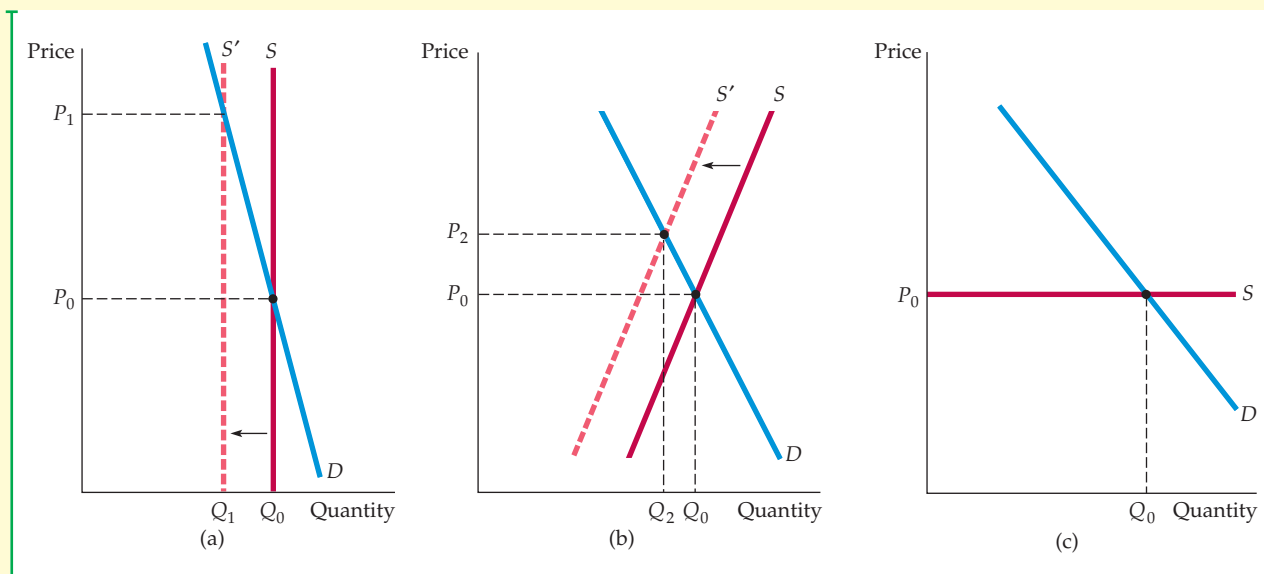


FIGURE 2.18
SUPPLY AND DEMAND FOR COFFEE

(a) A freeze or drought in Brazil causes the supply curve to shift to the left. In the short run, supply is completely inelastic; only a fixed number of coffee beans can be harvested. Demand is also relatively inelastic; consumers change their habits only slowly. As a result, the initial effect of the freeze is a sharp increase in price, from P_0 to P_1 . **(b)** In the intermediate run, supply and demand are both more elastic; thus price falls part of the way back, to P_2 . **(c)** In the long run, supply is extremely elastic; because new coffee trees will have had time to mature, the effect of the freeze will have disappeared. Price returns to P_0 .

In the intermediate run—say, one year after the freeze—both supply and demand are more elastic, supply because existing trees can be harvested more intensively (with some decrease in quality), and demand because consumers have had time to change their buying habits. As part (b) shows, although the intermediate-run supply curve also shifts to the left, price has come down from P_1 to P_2 .

The quantity supplied has also increased somewhat from the short run, from Q_1 to Q_2 . In the long run shown in part (c), price returns to its normal level because growers have had time to replace trees damaged by the freeze. The long-run supply curve, then, simply reflects the cost of producing coffee, including the costs of land, of planting and caring for the trees, and of a competitive rate of profit.¹⁴

*2.6 Understanding and Predicting the Effects of Changing Market Conditions

So far, our discussion of supply and demand has been largely qualitative. To use supply and demand curves to analyze and predict the effects of changing market conditions, we must begin attaching numbers to them. For example, to see how a 50-percent reduction in the supply of Brazilian coffee may affect the world price of coffee, we must determine actual supply and demand

¹⁴You can learn more about the world coffee market from the Foreign Agriculture Service of the U.S. Department of Agriculture by visiting their Web site at <http://www.fas.usda.gov/http/coffee.asp>. Another good source of information is <http://www.nationalgeographic.com/coffee>.



curves and then calculate the shifts in those curves and the resulting changes in price.

In this section, we will see how to do simple “back of the envelope” calculations with linear supply and demand curves. Although they are often approximations of more complex curves, we use linear curves because they are easier to work with. It may come as a surprise, but one can do some informative economic analyses on the back of a small envelope with a pencil and a pocket calculator.

First, we must learn how to “fit” linear demand and supply curves to market data. (By this we do not mean *statistical fitting* in the sense of linear regression or other statistical techniques, which we will discuss later in the book.) Suppose we have two sets of numbers for a particular market: The first set consists of the price and quantity that generally prevail in the market (i.e., the price and quantity that prevail “on average,” when the market is in equilibrium or when market conditions are “normal”). We call these numbers the *equilibrium price* and *quantity* and denote them by P^* and Q^* . The second set consists of the price elasticities of supply and demand for the market (at or near the equilibrium), which we denote by E_s and E_D , as before.

These numbers may come from a statistical study done by someone else; they may be numbers that we simply think are reasonable; or they may be numbers that we want to try out on a “what if” basis. Our goal is to *write down the supply and demand curves that fit (i.e., are consistent with) these numbers*. We can then determine numerically how a change in a variable such as GDP, the price of another good, or some cost of production will cause supply or demand to shift and thereby affect market price and quantity.

Let’s begin with the linear curves shown in Figure 2.19. We can write these curves algebraically as follows:

$$\text{Demand: } Q = a - bP \quad (2.5a)$$

$$\text{Supply: } Q = c + dP \quad (2.5b)$$

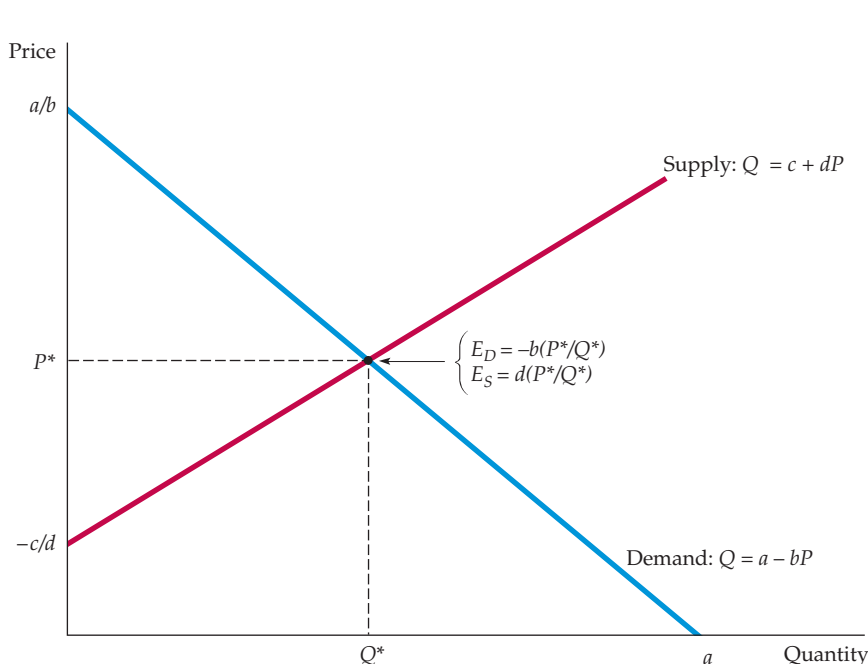


FIGURE 2.19
FITTING LINEAR SUPPLY AND DEMAND CURVES TO DATA

Linear supply and demand curves provide a convenient tool for analysis. Given data for the equilibrium price and quantity P^* and Q^* , as well as estimates of the elasticities of demand and supply E_D and E_S , we can calculate the parameters c and d for the supply curve and a and b for the demand curve. (In the case drawn here, $c < 0$.) The curves can then be used to analyze the behavior of the market quantitatively.



Our problem is to choose numbers for the constants a , b , c , and d . This is done, for supply and for demand, in a two-step procedure:

- **Step 1:** Recall that each price elasticity, whether of supply or demand, can be written as

$$E = (P/Q)(\Delta Q/\Delta P)$$

where $\Delta Q/\Delta P$ is the change in quantity demanded or supplied resulting from a small change in price. For linear curves, $\Delta Q/\Delta P$ is constant. From equations (2.5a) and (2.5b), we see that $\Delta Q/\Delta P = d$ for supply and $\Delta Q/\Delta P = -b$ for demand. Now, let's substitute these values for $\Delta Q/\Delta P$ into the elasticity formula:

$$\text{Demand: } E_D = -b(P^*/Q^*) \quad (2.6a)$$

$$\text{Supply: } E_S = d(P^*/Q^*) \quad (2.6b)$$

where P^* and Q^* are the equilibrium price and quantity for which we have data and to which we want to fit the curves. Because we have numbers for E_S , E_D , P^* , and Q^* , we can substitute these numbers in equations (2.6a) and (2.6b) and solve for b and d .

- **Step 2:** Since we now know b and d , we can substitute these numbers, as well as P^* and Q^* , into equations (2.5a) and (2.5b) and solve for the remaining constants a and c . For example, we can rewrite equation (2.5a) as

$$a = Q^* + bP^*$$

and then use our data for Q^* and P^* , together with the number we calculated in Step 1 for b , to obtain a .

Let's apply this procedure to a specific example: long-run supply and demand for the world copper market. The relevant numbers for this market are as follows:

Quantity $Q^* = 18$ million metric tons per year (mmt/yr)

Price $P^* = \$3.00$ per pound

Elasticity of supply $E_S = 1.5$

Elasticity of demand $E_D = -0.5$.

(The price of copper has fluctuated during the past few decades between \$0.60 and more than \$4.00, but \$3.00 is a reasonable average price for 2008–2011).

We begin with the supply curve equation (2.5b) and use our two-step procedure to calculate numbers for c and d . The long-run price elasticity of supply is 1.5, $P^* = \$3.00$, and $Q^* = 18$.

- **Step 1:** Substitute these numbers in equation (2.6b) to determine d :

$$1.5 = d(3/18) = d/6$$

so that $d = (1.5)(6) = 9$.

- **Step 2:** Substitute this number for d , together with the numbers for P^* and Q^* , into equation (2.5b) to determine c :

$$18 = c + (9)(3.00) = c + 27$$



so that $c = 18 - 27 = -9$. We now know c and d , so we can write our supply curve:

$$\text{Supply: } Q = -9 + 9P$$

We can now follow the same steps for the demand curve equation (2.5a). An estimate for the long-run elasticity of demand is -0.5 .¹⁵ First, substitute this number, as well as the values for P^* and Q^* , into equation (2.6a) to determine b :

$$-0.5 = -b(3/18) = -b/6$$

so that $b = (0.5)(6) = 3$. Second, substitute this value for b and the values for P^* and Q^* in equation (2.5a) to determine a :

$$18 = a = (3)(3) = a - 9$$

so that $a = 18 + 9 = 27$. Thus, our demand curve is:

$$\text{Demand: } Q = 27 - 3P$$

To check that we have not made a mistake, let's set the quantity supplied equal to the quantity demanded and calculate the resulting equilibrium price:

$$\begin{aligned} \text{Supply} &= -9 + 9P = 27 - 3P = \text{Demand} \\ 9P + 3P &= 27 + 9 \end{aligned}$$

or $P = 36/12 = 3.00$, which is indeed the equilibrium price with which we began.

Although we have written supply and demand so that they depend only on price, they could easily depend on other variables as well. Demand, for example, might depend on income as well as price. We would then write demand as

$$Q = a - bP + fI \quad (2.7)$$

where I is an index of the aggregate income or GDP. For example, I might equal 1.0 in a base year and then rise or fall to reflect percentage increases or decreases in aggregate income.

For our copper market example, a reasonable estimate for the long-run income elasticity of demand is 1.3. For the linear demand curve (2.7), we can then calculate f by using the formula for the income elasticity of demand: $E = (I/Q)(\Delta Q/\Delta I)$. Taking the base value of I as 1.0, we have

$$1.3 = (1.0/18)(f).$$

Thus $f = (1.3)(18)/(1.0) = 23.4$. Finally, substituting the values $b = 3$, $f = 23.4$, $P^* = 3.00$, and $Q^* = 18$ into equation (2.7), we can calculate that a must equal 3.6.

¹⁵See Claudio Agostini, "Estimating Market Power in the U.S. Copper Industry," *Review of Industrial Organization* 28 (2006), 17–39.



We have seen how to fit linear supply and demand curves to data. Now, to see how these curves can be used to analyze markets, let's look at Example 2.8, which deals with the behavior of copper prices, and Example 2.9, which concerns the world oil market.

EXAMPLE 2.8 THE BEHAVIOR OF COPPER PRICES

After reaching a level of about \$1.00 per pound in 1980, the price of copper fell sharply to about 60 cents per pound in 1986. In real (inflation-adjusted) terms, this price was even lower than during the Great Depression 50 years earlier. Prices increased in 1988–1989 and in 1995, largely as a result of strikes by miners in Peru and Canada that disrupted supplies, but then fell again from 1996 through 2003. Prices increased sharply, however, between 2003 and 2007, and while copper fell along with many other commodities during the 2008–2009 recession,

the price of copper had recovered by early 2010. Figure 2.20 shows the behavior of copper prices from 1965 to 2011 in both real and nominal terms.

Worldwide recessions in 1980 and 1982 contributed to the decline of copper prices; as mentioned above, the income elasticity of copper demand is about 1.3. But copper demand did not pick up as the industrial economies recovered during the mid-1980s. Instead, the 1980s saw a steep decline in demand.

The price decline through 2003 occurred for two reasons. First, a large part of copper consumption is

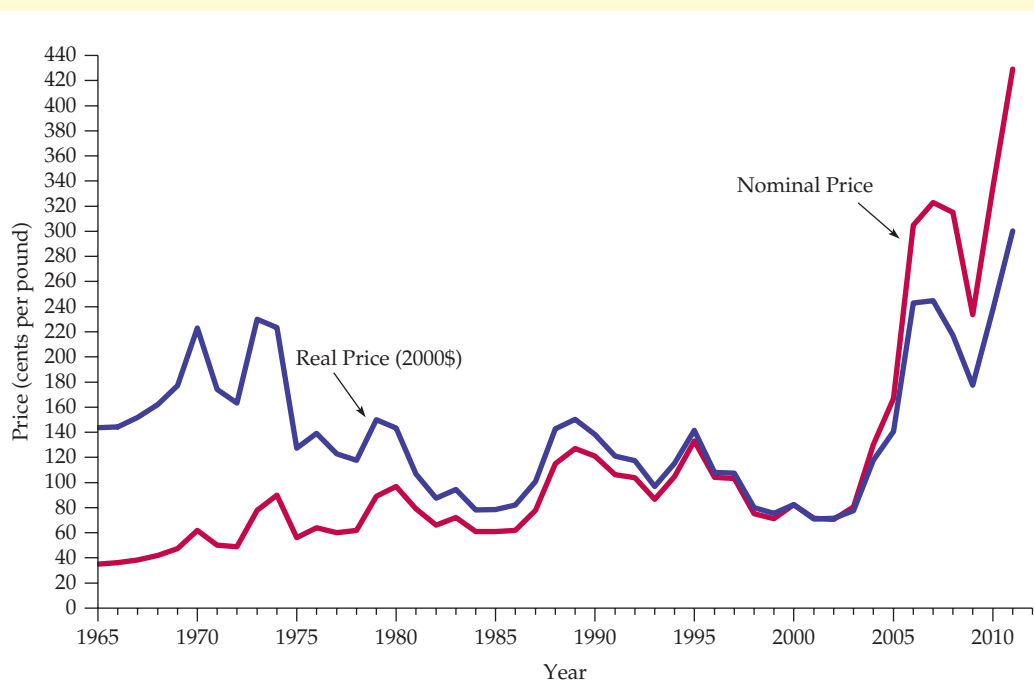


FIGURE 2.20
COPPER PRICES, 1965–2011

Copper prices are shown in both nominal (no adjustment for inflation) and real (inflation-adjusted) terms. In real terms, copper prices declined steeply from the early 1970s through the mid-1980s as demand fell. In 1988–1990, copper prices rose in response to supply disruptions caused by strikes in Peru and Canada but later fell after the strikes ended. Prices declined during the 1996–2002 period but then increased sharply starting in 2005.



for the construction of equipment for electric power generation and transmission. But by the late 1970s, the growth rate of electric power generation had fallen dramatically in most industrialized countries. In the United States, for example, the growth rate fell from over 6 percent per annum in the 1960s and early 1970s to less than 2 percent in the late 1970s and 1980s. This decline meant a big drop in what had been a major source of copper demand. Second, in the 1980s, other materials, such as aluminum and fiber optics, were increasingly substituted for copper.

Why did the price increase so sharply after 2003? First, the demand for copper from China and other Asian countries began increasing dramatically, replacing the demand from Europe and the U.S.

Chinese copper consumption, for example, has nearly tripled since 2001. Second, because prices had dropped so much from 1996 through 2003, producers in the U.S., Canada, and Chile closed unprofitable mines and cut production. Between 2000 and 2003, for example, U.S. mine production of copper declined by 23 percent.¹⁶

One might expect increasing prices to stimulate investments in new mines and increases in production, and that is indeed what has happened. Arizona, for example, experienced a copper boom as Phelps Dodge opened a major new mine in 2007.¹⁷ By 2007, producers began to worry that prices would decline again, either as a result of these new investments or because demand from Asia would level off or even drop.

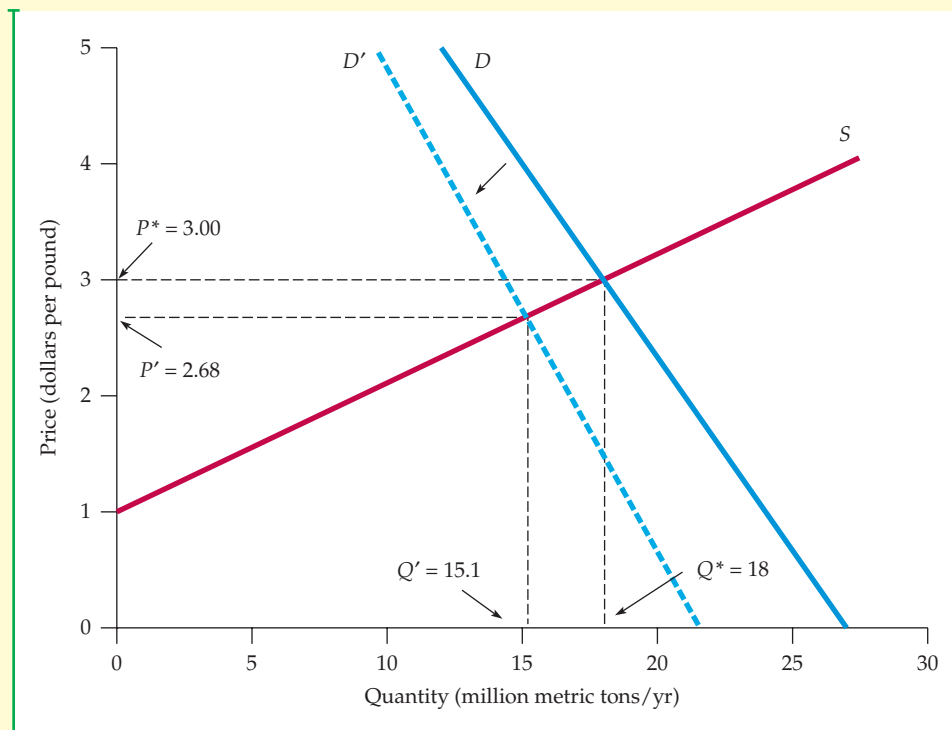


FIGURE 2.21
COPPER SUPPLY AND DEMAND

The shift in the demand curve corresponding to a 20-percent decline in demand leads to a 10.7-percent decline in price.

¹⁶Our thanks to Patricia Foley, Executive Director of the American Bureau of Metal Statistics, for supplying the data on China. Other data are from the Monthly Reports of the U.S. Geological Survey Mineral Resources Program—<http://minerals.usgs.gov/minerals/pubs/copper>.

¹⁷The boom created hundreds of new jobs, which in turn led to increases in housing prices: “Copper Boom Creates Housing Crunch,” *The Arizona Republic*, July 12, 2007.



What would a decline in demand do to the price of copper? To find out, we can use the linear supply and demand curves that we just derived. Let's calculate the effect on price of a 20-percent decline in demand. Because we are not concerned here with the effects of GDP growth, we can leave the income term, I , out of the demand equation.

We want to shift the demand curve to the left by 20 percent. In other words, we want the quantity demanded to be 80 percent of what it would be otherwise for every value of price. For our linear demand curve, we simply multiply the right-hand side by 0.8:

$$Q = (0.8)(27 - 3P) = 21.6 - 2.4P$$

Supply is again $Q = -9 + 9P$. Now we can equate the quantity supplied and the quantity demanded and solve for price:

$$-9 + 9P = 21.6 - 2.4P$$

or $P = 30.6/11.4 = \$2.68$ per pound. A decline in demand of 20 percent, therefore, entails a drop in price of roughly 32 cents per pound, or 10.7 percent.¹⁸

EXAMPLE 2.9 UPHEAVAL IN THE WORLD OIL MARKET



Since the early 1970s, the world oil market has been buffeted by the OPEC cartel and by political turmoil in the Persian Gulf. In 1974, by collectively restraining output, OPEC (the Organization of Petroleum Exporting Countries) pushed world oil prices well above what they would have been in a competitive market. OPEC could do this because it accounted for much of world oil production. During

1979–1980, oil prices shot up again, as the Iranian revolution and the outbreak of the Iran-Iraq war sharply reduced Iranian and Iraqi production. During the 1980s, the price gradually declined, as demand fell and competitive (i.e., non-OPEC) supply rose in response to price. Prices remained relatively stable during 1988–2001, except for a temporary spike in 1990 following the Iraqi invasion of Kuwait. Prices increased again in 2002–2003 as a result of a strike in Venezuela and then the war with Iraq that began in the spring of 2003. Oil prices continued to increase through the summer of 2008 as a result of rising demand in Asia and reductions in OPEC output. By the end of 2008, the recession had reduced demand around the world, leading prices to plummet 127% in six months. Between 2009 and 2011, oil prices have gradually recovered, partially buoyed by China's continuing growth. Figure 2.22 shows the world price of oil from 1970 to 2011, in both nominal and real terms.¹⁹

The Persian Gulf is one of the less stable regions of the world—a fact that has led to concern over the possibility of new oil supply disruptions and sharp increases in oil prices. What would happen to oil prices—in both the

¹⁸Note that because we have multiplied the demand function by 0.8—i.e., reduced the quantity demanded at every price by 20 percent—the new demand curve is not parallel to the old one. Instead, the curve rotates downward at its intersection with the price axis.

¹⁹For a nice overview of the factors that have affected world oil prices, see James D. Hamilton, “Understanding Crude Oil Prices,” *The Energy Journal*, 2009, Vol. 30, pp. 179–206.

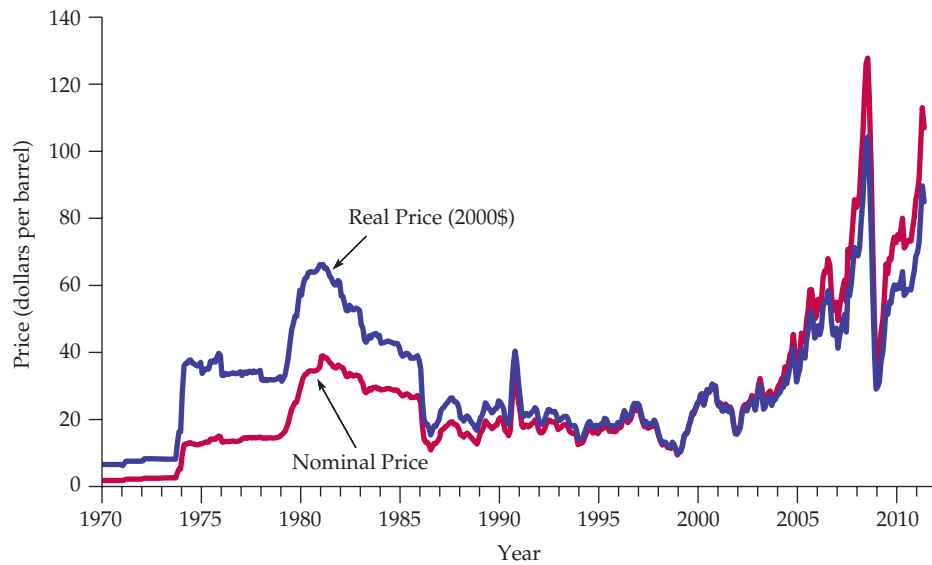


FIGURE 2.22
PRICE OF CRUDE OIL

The OPEC cartel and political events caused the price of oil to rise sharply at times. It later fell as supply and demand adjusted.

short run and longer run—if a war or revolution in the Persian Gulf caused a sharp cutback in oil production? Let's see how simple supply and demand curves can be used to predict the outcome of such an event.

Because this example is set in 2009–2011, all prices are measured in 2011 dollars. Here are some rough figures:

- 2009–2011 world price = \$80 per barrel
- World demand and total supply = 32 billion barrels per year (bb/yr)
- OPEC supply = 13 bb/yr
- Competitive (non-OPEC) supply = 19 bb/yr

The following table gives price elasticity estimates for oil supply and demand:²⁰

	SHORT RUN	LONG RUN
World demand:	−0.05	−0.30
Competitive supply:	0.05	0.30

²⁰For the sources of these numbers and a more detailed discussion of OPEC oil pricing, see Robert S. Pindyck, "Gains to Producers from the Cartelization of Exhaustible Resources," *Review of Economics and Statistics* 60 (May 1978): 238–51; James M. Griffin and David J. Teece, *OPEC Behavior and World Oil Prices* (London: Allen and Unwin, 1982); and John C. B. Cooper, "Price Elasticity of Demand for Crude Oil: Estimates for 23 Countries," *Organization of the Petroleum Exporting Countries Review* (March 2003).



You should verify that these numbers imply the following for demand and competitive supply in the *short run*:

$$\text{Short-run demand: } D = 33.6 - .020P$$

$$\text{Short-run competitive supply: } S_C = 18.05 + 0.012P$$

Of course, *total supply* is competitive supply *plus* OPEC supply, which we take as constant at 13 bb/yr. Adding this 13 bb/yr to the competitive supply curve above, we obtain the following for the total short-run supply:

$$\text{Short-run total supply: } S_T = 31.05 + 0.012P$$

You should verify that the quantity demanded and the total quantity supplied are equal at an equilibrium price of \$80 per barrel.

You should also verify that the corresponding demand and supply curves for the *long run* are as follows:

$$\text{Long-run demand: } D = 41.6 - 0.120P$$

$$\text{Long-run competitive supply: } S_C = 13.3 + 0.071P$$

$$\text{Long-run total supply: } S_T = 26.3 + 0.071P$$

Again, you can check that the quantities supplied and demanded equate at a price of \$80.

Saudi Arabia is one of the world's largest oil producers, accounting for roughly 3 bb/yr, which is nearly 10 percent of total world production. What would happen to the price of oil if, because of war or political upheaval, Saudi Arabia stopped producing oil? We can use our supply and demand curves to find out.

For the *short run*, simply subtract 3 from short-run total supply:

$$\text{Short-run demand: } D = 33.6 - .020P$$

$$\text{Short-run total supply: } S_T = 28.05 + 0.012P$$

By equating this total quantity supplied with the quantity demanded, we can see that in the short run, the price will more than double to \$173.44 per barrel. Figure 2.23 shows this supply shift and the resulting short-run increase in price. The initial equilibrium is at the intersection of S_T and D . After the drop in Saudi production, the equilibrium occurs where S_T' and D cross.

In the *long run*, however, things will be different. Because both demand and competitive supply are more elastic in the long run, the 3 bb/yr cut in oil production will no longer support such a high price. Subtracting 3 from long-run total supply and equating with long-run demand, we can see that the price will fall to \$95.81, only \$15.81 above the initial \$80 price.

Thus, if Saudi Arabia suddenly stops producing oil, we should expect to see about a doubling in price. However, we should also expect to see the price gradually decline afterward, as demand falls and competitive supply rises.

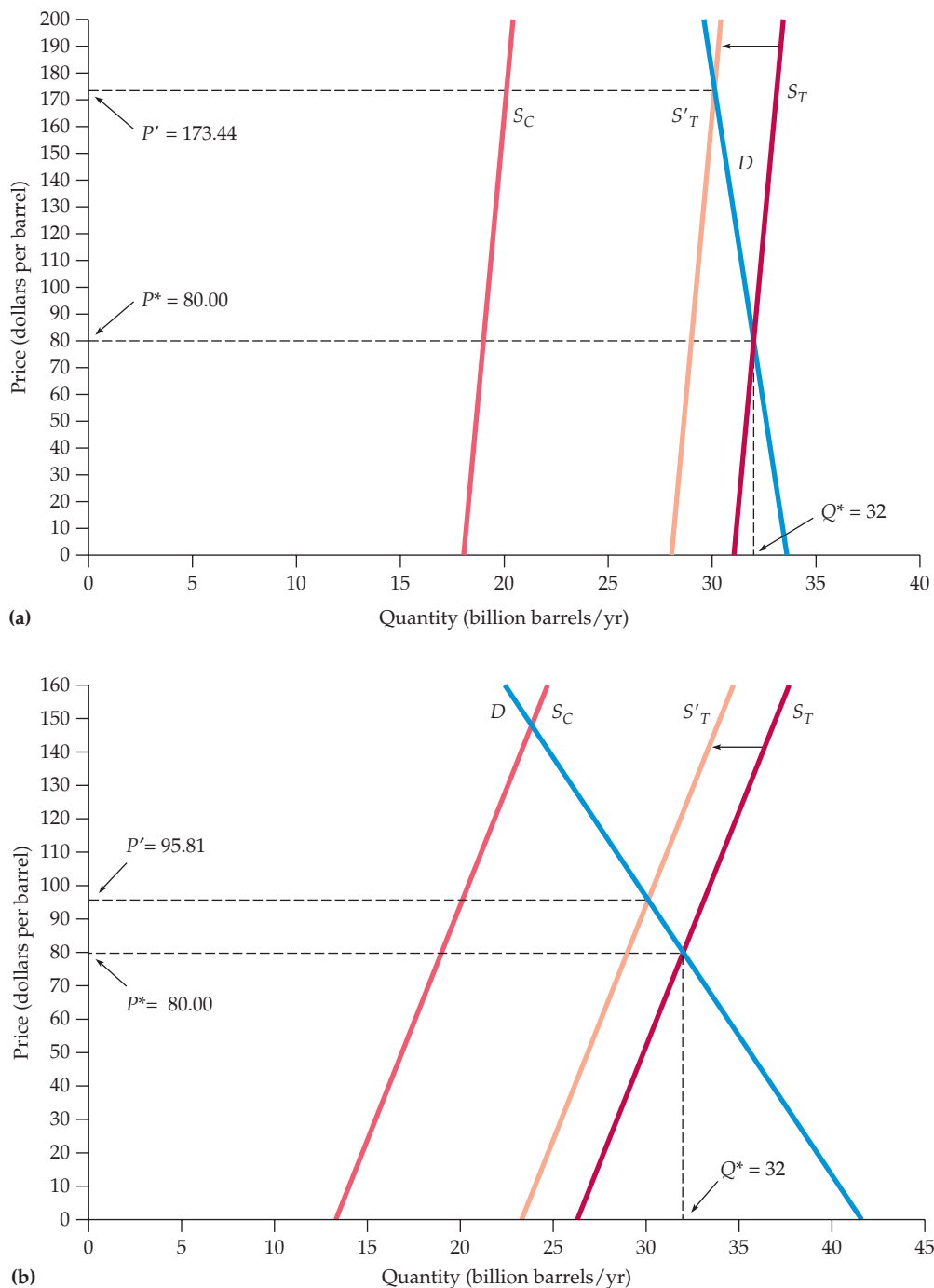


FIGURE 2.23
IMPACT OF SAUDI PRODUCTION CUT

The total supply is the sum of competitive (non-OPEC) supply and the 13 bb/yr of OPEC supply. Part (a) shows the short-run supply and demand curves. If Saudi Arabia stops producing, the supply curve will shift to the left by 3 bb/yr. In the short-run, price will increase sharply. Part (b) shows long-run curves. In the long run, because demand and competitive supply are much more elastic, the impact on price will be much smaller.



This is indeed what happened following the sharp decline in Iranian and Iraqi production in 1979–1980. History may or may not repeat itself, but if it does, we can at least predict the impact on oil prices.²¹

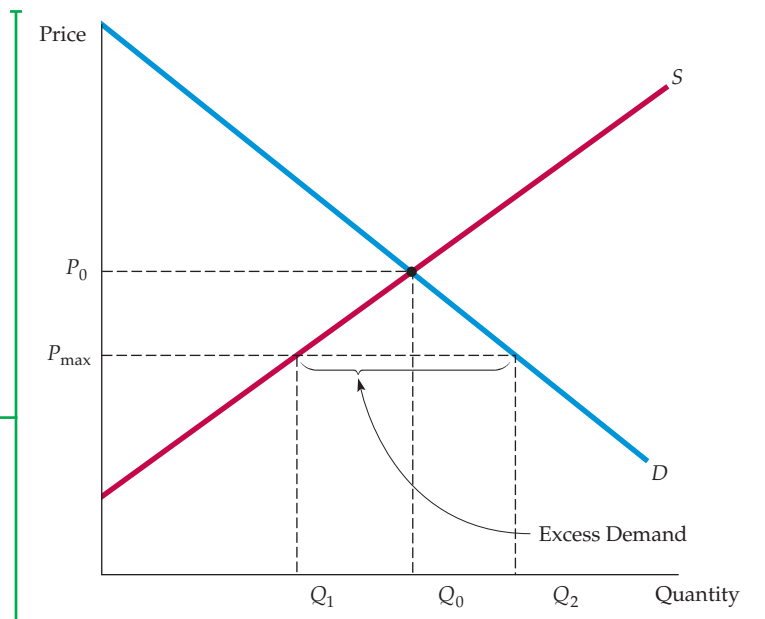
2.7 Effects of Government Intervention—Price Controls

In the United States and most other industrial countries, markets are rarely free of government intervention. Besides imposing taxes and granting subsidies, governments often regulate markets (even competitive markets) in a variety of ways. In this section, we will see how to use supply and demand curves to analyze the effects of one common form of government intervention: price controls. Later, in Chapter 9, we will examine the effects of price controls and other forms of government intervention and regulation in more detail.

Figure 2.24 illustrates the effects of price controls. Here, P_0 and Q_0 are the equilibrium price and quantity that would prevail without government regulation. The government, however, has decided that P_0 is too high and mandated that the price can be no higher than a maximum allowable *ceiling price*, denoted by P_{\max} . What is the result? At this lower price, producers (particularly those with higher costs) will produce less, and the quantity supplied will drop to Q_1 . Consumers, on the other hand, will demand more at this low price; they would like to purchase the quantity Q_2 . Demand therefore exceeds supply, and a shortage develops—i.e., there is *excess demand*. The amount of excess demand is $Q_2 - Q_1$.

FIGURE 2.24
EFFECTS OF PRICE CONTROLS

Without price controls, the market clears at the equilibrium price and quantity P_0 and Q_0 . If price is regulated to be no higher than P_{\max} , the quantity supplied falls to Q_1 , the quantity demanded increases to Q_2 , and a shortage develops.



²¹You can obtain recent data and learn more about the world oil market by accessing the Web sites of the American Petroleum Institute at www.api.org or the U.S. Energy Information Administration at www.eia.doe.gov.



This excess demand sometimes takes the form of queues, as when drivers lined up to buy gasoline during the winter of 1974 and the summer of 1979. In both instances, the lines were the result of price controls; the government prevented domestic oil and gasoline prices from rising along with world oil prices. Sometimes excess demand results in curtailments and supply rationing, as with natural gas price controls and the resulting gas shortages of the mid-1970s, when industrial consumers closed factories because gas supplies were cut off. Sometimes it spills over into other markets, where it artificially increases demand. For example, natural gas price controls caused potential buyers of gas to use oil instead.

Some people gain and some lose from price controls. As Figure 2.24 suggests, producers lose: They receive lower prices, and some leave the industry. Some but not all consumers gain. While those who can purchase the good at a lower price are better off, those who have been “rationed out” and cannot buy the good at all are worse off. How large are the gains to the winners and how large are the losses to the losers? Do total gains exceed total losses? To answer these questions, we need a method to measure the gains and losses from price controls and other forms of government intervention. We discuss such a method in Chapter 9.

EXAMPLE 2.10 PRICE CONTROLS AND NATURAL GAS SHORTAGES

In 1954, the federal government began regulating the wellhead price of natural gas. Initially the controls were not binding; the ceiling prices were above those that cleared the market. But in about 1962, when these ceiling prices did become binding, excess demand for natural gas developed and slowly began to grow. In the 1970s, this excess demand, spurred by higher oil prices, became severe and led to widespread curtailments. Soon ceiling prices were far below prices that would have prevailed in a free market.²²

Today, producers and industrial consumers of natural gas, oil, and other commodities are concerned that the government might respond, once again, with price controls if prices rise sharply. Let's calculate the likely impact of price controls on natural gas, based on market conditions in 2007.

Figure 2.25 shows the wholesale price of natural gas, in both nominal and real (2000 dollars) terms, from 1950 through 2007. The following numbers describe the U.S. market in 2007:

- The (free-market) wholesale price of natural gas was \$6.40 per mcf (thousand cubic feet);
- Production and consumption of gas were 23 Tcf (trillion cubic feet);
- The average price of crude oil (which affects the supply and demand for natural gas) was about \$50 per barrel.

A reasonable estimate for the price elasticity of supply is 0.2. Higher oil prices also lead to more natural gas production because oil and gas are often discovered and produced together; an estimate of the cross-price elasticity of supply is 0.1. As for demand, the price elasticity is about -0.5, and the cross-price elasticity with respect to oil price is about 1.5. You can verify that the following linear supply and demand curves fit these numbers:

$$\text{Supply: } Q = 15.90 + 0.72P_G + 0.05P_O$$

$$\text{Demand: } Q = 0.02 - 1.8P_G + 0.69P_O$$

²²This regulation began with the Supreme Court's 1954 decision requiring the then Federal Power Commission to regulate wellhead prices on natural gas sold to interstate pipeline companies. These price controls were largely removed during the 1980s, under the mandate of the Natural Gas Policy Act of 1978. For a detailed discussion of natural gas regulation and its effects, see Paul W. MacAvoy and Robert S. Pindyck, *The Economics of the Natural Gas Shortage* (Amsterdam: North-Holland, 1975); R. S. Pindyck, “Higher Energy Prices and the Supply of Natural Gas,” *Energy Systems and Policy* 2(1978): 177–209; and Arlon R. Tussing and Connie C. Barlow, *The Natural Gas Industry* (Cambridge, MA: Ballinger, 1984).

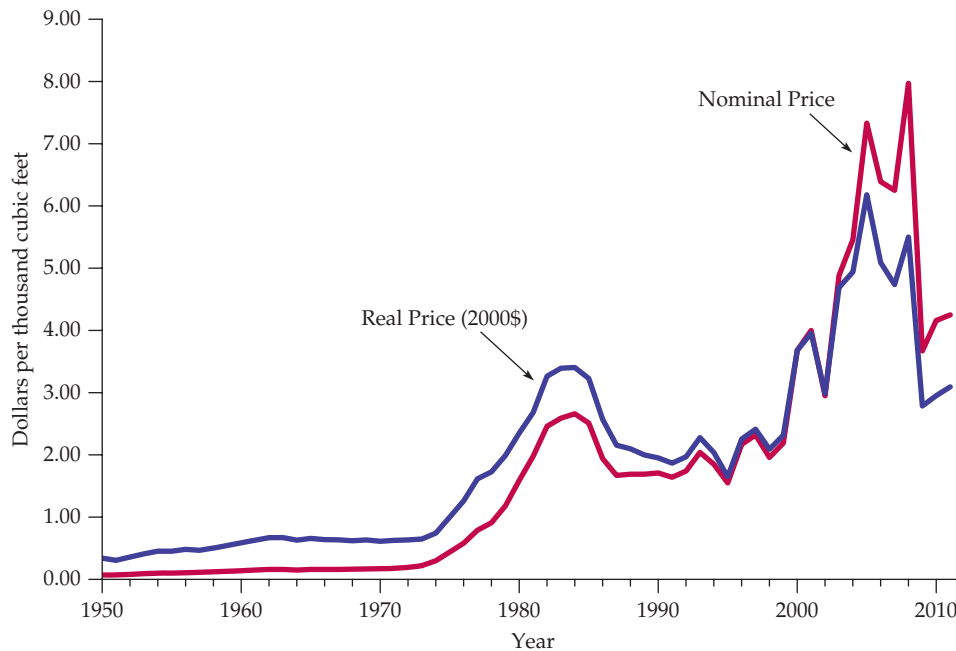


FIGURE 2.25
PRICE OF NATURAL GAS

Natural gas prices rose sharply after 2000, as did the prices of oil and other fuels.

where Q is the quantity of natural gas (in Tcf), P_G is the price of natural gas (in dollars per mcf), and P_O is the price of oil (in dollars per barrel). You can also verify, by equating the quantities supplied and demanded and substituting \$50 for P_O , that these supply and demand curves imply an equilibrium free-market price of \$6.40 for natural gas.

Suppose the government determines that the free-market price of \$6.40 per mcf is too high, decides to impose price controls, and sets a maximum price of \$3.00 per mcf. What impact would

this have on the quantity of gas supplied and the quantity demanded?

Substitute \$3.00 for P_G in both the supply and demand equations (keeping the price of oil, P_O , fixed at \$50). You should find that the supply equation gives a quantity supplied of 20.6 Tcf and the demand equation a quantity demanded of 29.1 Tcf. Therefore, these price controls would create an excess demand (i.e., shortage) of $29.1 - 20.6 = 8.5$ Tcf. In Example 9.1 we'll show how to measure the resulting gains and losses to producers and consumers.

SUMMARY

1. Supply-demand analysis is a basic tool of microeconomics. In competitive markets, supply and demand curves tell us how much will be produced by firms and how much will be demanded by consumers as a function of price.
2. The market mechanism is the tendency for supply and demand to equilibrate (i.e., for price to move to the

market-clearing level), so that there is neither excess demand nor excess supply. The equilibrium price is the price that equates the quantity demanded with the quantity supplied.

3. Elasticities describe the responsiveness of supply and demand to changes in price, income, or other variables. For example, the price elasticity of demand measures



the percentage change in the quantity demanded resulting from a 1-percent increase in price.

4. Elasticities pertain to a time frame, and for most goods it is important to distinguish between short-run and long-run elasticities.
5. We can use supply-demand diagrams to see how shifts in the supply curve and/or demand curve can explain changes in the market price and quantity.
6. If we can estimate, at least roughly, the supply and demand curves for a particular market, we can calculate the market-clearing price by equating the quantity supplied with the quantity demanded. Also, if we know how supply and demand depend on other economic variables, such as income or the prices of other goods, we

can calculate how the market-clearing price and quantity will change as these other variables change. This is a means of explaining or predicting market behavior.

7. Simple numerical analyses can often be done by fitting linear supply and demand curves to data on price and quantity and to estimates of elasticities. For many markets, such data and estimates are available, and simple “back of the envelope” calculations can help us understand the characteristics and behavior of the market.
8. When a government imposes price controls, it keeps the price below the level that equates supply and demand. A shortage develops; the quantity demanded exceeds the quantity supplied.

QUESTIONS FOR REVIEW

1. Suppose that unusually hot weather causes the demand curve for ice cream to shift to the right. Why will the price of ice cream rise to a new market-clearing level?
2. Use supply and demand curves to illustrate how each of the following events would affect the price of butter and the quantity of butter bought and sold: (a) an increase in the price of margarine; (b) an increase in the price of milk; (c) a decrease in average income levels.
3. If a 3-percent increase in the price of corn flakes causes a 6-percent decline in the quantity demanded, what is the elasticity of demand?
4. Explain the difference between a shift in the supply curve and a movement along the supply curve.
5. Explain why for many goods, the long-run price elasticity of supply is larger than the short-run elasticity.
6. Why do long-run elasticities of demand differ from short-run elasticities? Consider two goods: paper towels and televisions. Which is a durable good? Would you expect the price elasticity of demand for paper towels to be larger in the short run or in the long run? Why? What about the price elasticity of demand for televisions?
7. Are the following statements true or false? Explain your answers.
 - a. The elasticity of demand is the same as the slope of the demand curve.
 - b. The cross-price elasticity will always be positive.
 - c. The supply of apartments is more inelastic in the short run than the long run.
8. Suppose the government regulates the prices of beef and chicken and sets them below their market-clearing levels. Explain why shortages of these goods will develop and what factors will determine the sizes of the shortages. What will happen to the price of pork? Explain briefly.
9. The city council of a small college town decides to regulate rents in order to reduce student living expenses. Suppose the average annual market-clearing rent for a two-bedroom apartment had been \$700 per month

and that rents were expected to increase to \$900 within a year. The city council limits rents to their current \$700-per-month level.

- a. Draw a supply and demand graph to illustrate what will happen to the rental price of an apartment after the imposition of rent controls.
- b. Do you think this policy will benefit all students? Why or why not?
10. In a discussion of tuition rates, a university official argues that the demand for admission is completely price inelastic. As evidence, she notes that while the university has doubled its tuition (in real terms) over the past 15 years, neither the number nor quality of students applying has decreased. Would you accept this argument? Explain briefly. (*Hint:* The official makes an assertion about the demand for admission, but does she actually observe a demand curve? What else could be going on?)
11. Suppose the demand curve for a product is given by

$$Q = 10 - 2P + P_s$$

where P is the price of the product and P_s is the price of a substitute good. The price of the substitute good is \$2.00.

- a. Suppose $P = \$1.00$. What is the price elasticity of demand? What is the cross-price elasticity of demand?
- b. Suppose the price of the good, P , goes to \$2.00. Now what is the price elasticity of demand? What is the cross-price elasticity of demand?
12. Suppose that rather than the declining demand assumed in Example 2.8, a decrease in the cost of copper production causes the supply curve to shift to the right by 40 percent. How will the price of copper change?
13. Suppose the demand for natural gas is perfectly inelastic. What would be the effect, if any, of natural gas price controls?



EXERCISES

- Suppose the demand curve for a product is given by $Q = 300 - 2P + 4I$, where I is average income measured in thousands of dollars. The supply curve is $Q = 3P - 50$.
 - If $I = 25$, find the market-clearing price and quantity for the product.
 - If $I = 50$, find the market-clearing price and quantity for the product.
 - Draw a graph to illustrate your answers.
- Consider a competitive market for which the quantities demanded and supplied (per year) at various prices are given as follows:

PRICE (DOLLARS)	DEMAND (MILLIONS)	SUPPLY (MILLIONS)
60	22	14
80	20	16
100	18	18
120	16	20

- Calculate the price elasticity of demand when the price is \$80 and when the price is \$100.
 - Calculate the price elasticity of supply when the price is \$80 and when the price is \$100.
 - What are the equilibrium price and quantity?
 - Suppose the government sets a price ceiling of \$80. Will there be a shortage, and if so, how large will it be?
- Refer to Example 2.5 (page 37) on the market for wheat. In 1998, the total demand for U.S. wheat was $Q = 3244 - 283P$ and the domestic supply was $Q_S = 1944 + 207P$. At the end of 1998, both Brazil and Indonesia opened their wheat markets to U.S. farmers. Suppose that these new markets add 200 million bushels to U.S. wheat demand. What will be the free-market price of wheat and what quantity will be produced and sold by U.S. farmers?
 - A vegetable fiber is traded in a competitive world market, and the world price is \$9 per pound. Unlimited quantities are available for import into the United States at this price. The U.S. domestic supply and demand for various price levels are shown as follows:

PRICE	U.S. SUPPLY (MILLION LBS)	U.S. DEMAND (MILLION LBS)
3	2	34
6	4	28
9	6	22
12	8	16
15	10	10
18	12	4

- What is the equation for demand? What is the equation for supply?
 - At a price of \$9, what is the price elasticity of demand? What is it at a price of \$12?
 - What is the price elasticity of supply at \$9? At \$12?
 - In a free market, what will be the U.S. price and level of fiber imports?
- Much of the demand for U.S. agricultural output has come from other countries. In 1998, the total demand for wheat was $Q = 3244 - 283P$. Of this, total domestic demand was $Q_D = 1700 - 107P$, and domestic supply was $Q_S = 1944 + 207P$. Suppose the export demand for wheat falls by 40 percent.
 - U.S. farmers are concerned about this drop in export demand. What happens to the free-market price of wheat in the United States? Do farmers have much reason to worry?
 - Now suppose the U.S. government wants to buy enough wheat to raise the price to \$3.50 per bushel. With the drop in export demand, how much wheat would the government have to buy? How much would this cost the government?
 - The rent control agency of New York City has found that aggregate demand is $Q_D = 160 - 8P$. Quantity is measured in tens of thousands of apartments. Price, the average monthly rental rate, is measured in hundreds of dollars. The agency also noted that the increase in Q at lower P results from more three-person families coming into the city from Long Island and demanding apartments. The city's board of realtors acknowledges that this is a good demand estimate and has shown that supply is $Q_S = 70 + 7P$.
 - If both the agency and the board are right about demand and supply, what is the free-market price? What is the change in city population if the agency sets a maximum average monthly rent of \$300 and all those who cannot find an apartment leave the city?
 - Suppose the agency bows to the wishes of the board and sets a rental of \$900 per month on all apartments to allow landlords a "fair" rate of return. If 50 percent of any long-run increases in apartment offerings comes from new construction, how many apartments are constructed?
 - In 2010, Americans smoked 315 billion cigarettes, or 15.75 billion packs of cigarettes. The average retail price (including taxes) was about \$5.00 per pack. Statistical studies have shown that the price elasticity of demand is -0.4 , and the price elasticity of supply is 0.5 .
 - Using this information, derive linear demand and supply curves for the cigarette market.
 - In 1998, Americans smoked 23.5 billion packs of cigarettes, and the retail price was about \$2.00 per pack. The decline in cigarette consumption from 1998 to 2010 was due in part to greater public awareness of the health hazards from smoking, but was also due in part to the increase in price. Suppose that the *entire decline* was due to the



increase in price. What could you deduce from that about the price elasticity of demand?

8. In Example 2.8 we examined the effect of a 20-percent decline in copper demand on the price of copper, using the linear supply and demand curves developed in Section 2.6. Suppose the long-run price elasticity of copper demand were -0.75 instead of -0.5 .
 - a. Assuming, as before, that the equilibrium price and quantity are $P^* = \$3$ per pound and $Q^* = 18$ million metric tons per year, derive the linear demand curve consistent with the smaller elasticity.
 - b. Using this demand curve, recalculate the effect of a 20-percent decline in copper demand on the price of copper.
9. In Example 2.8 (page 52), we discussed the recent increase in world demand for copper, due in part to China's rising consumption.
 - a. Using the original elasticities of demand and supply (i.e., $E_S = 1.5$ and $E_D = -0.5$), calculate the effect of a 20-percent *increase* in copper demand on the price of copper.
 - b. Now calculate the effect of this increase in demand on the equilibrium quantity, Q^* .
 - c. As we discussed in Example 2.8, the U.S. production of copper declined between 2000 and 2003. Calculate the effect on the equilibrium price and quantity of *both* a 20-percent increase in copper demand (as you just did in part a) *and* of a 20-percent decline in copper supply.
10. Example 2.9 (page 54) analyzes the world oil market. Using the data given in that example:
 - a. Show that the short-run demand and competitive supply curves are indeed given by

$$D = 33.6 - .020P$$

$$S_C = 18.05 + 0.012P$$

- b. Show that the long-run demand and competitive supply curves are indeed given by

$$D = 41.6 - 0.120P$$

$$S_C = 13.3 + 0.071P$$

- c. In Example 2.9 we examined the impact on price of a disruption of oil from Saudi Arabia. Suppose

that instead of a decline in supply, OPEC production *increases* by 2 billion barrels per year (bb/yr) because the Saudis open large new oil fields. Calculate the effect of this increase in production on the price of oil in both the short run and the long run.

11. Refer to Example 2.10 (page 59), which analyzes the effects of price controls on natural gas.
 - a. Using the data in the example, show that the following supply and demand curves describe the market for natural gas in 2005–2007:

$$\text{Supply: } Q = 15.90 + 0.72P_G + 0.05P_O$$

$$\text{Demand: } Q = 0.02 - 1.8P_G + 0.69P_O$$

Also, verify that if the price of oil is \$50, these curves imply a free-market price of \$6.40 for natural gas.

- b. Suppose the regulated price of gas were \$4.50 per thousand cubic feet instead of \$3.00. How much excess demand would there have been?
 - c. Suppose that the market for natural gas remained unregulated. If the price of oil had increased from \$50 to \$100, what would have happened to the free-market price of natural gas?
- *12. The table below shows the retail price and sales for instant coffee and roasted coffee for two years.
- a. Using these data alone, estimate the short-run price elasticity of demand for roasted coffee. Derive a linear demand curve for roasted coffee.
 - b. Now estimate the short-run price elasticity of demand for instant coffee. Derive a linear demand curve for instant coffee.
 - c. Which coffee has the higher short-run price elasticity of demand? Why do you think this is the case?

YEAR	RETAIL PRICE OF INSTANT COFFEE (\$/LB)	SALES OF INSTANT COFFEE (MILLION LBS)	RETAIL PRICE OF ROASTED COFFEE (\$/LB)	SALES OF ROASTED COFFEE (MILLION LBS)
Year 1	10.35	75	4.11	820
Year 2	10.48	70	3.76	850

CHAPTER 3

The Analysis of Competitive Markets

In Chapter 2, we saw how supply and demand curves can help us describe and understand the behavior of competitive markets.

Building on this foundation, we return to supply–demand analysis and show how it can be applied to a wide variety of economic problems—problems that might concern a consumer faced with a purchasing decision, a firm faced with a long-range planning problem, or a government agency that has to design a policy and evaluate its likely impact.

We begin by showing how consumer and producer surplus can be used to study the *welfare effects* of a government policy—in other words, who gains and who loses from the policy, and by how much. We also use consumer and producer surplus to demonstrate the *efficiency* of a competitive market—why the equilibrium price and quantity in a competitive market maximizes the aggregate economic welfare of producers and consumers.

Then we apply supply–demand analysis to a variety of problems. Because very few markets in the United States have been untouched by government interventions of one kind or another, most of the problems that we will study deal with the effects of such interventions. Our objective is not simply to solve these problems, but to show you how to use the tools of economic analysis to deal with them and others like them on your own. We hope that by working through the examples we provide, you will see how to calculate the response of markets to changing economic conditions or government policies and to evaluate the resulting gains and losses to consumers and producers.

3.1 Evaluating the Gains and Losses from Government Policies—Consumer and Producer Surplus

We saw at the end of Chapter 2 that a government-imposed price ceiling causes the quantity of a good demanded to rise (at the lower price, consumers want to buy more) and the quantity supplied to fall (producers are not willing to supply as much at the lower price). The result



CHAPTER OUTLINE

- 3.1** Evaluating the Gains and Losses from Government Policies—Consumer and Producer Surplus 317
- 3.2** The Efficiency of a Competitive Market 323
- 3.3** Minimum Prices 328
- 3.4** Price Supports and Production Quotas 332
- 3.5** Import Quotas and Tariffs 340
- 3.6** The Impact of a Tax or Subsidy 345

LIST OF EXAMPLES

- 3.1** Price Controls and Natural Gas Shortages 322
- 3.2** The Market for Human Kidneys 325
- 3.3** Airline Regulation 330
- 3.4** Supporting the Price of Wheat 335
- 3.5** Why Can't I Find a Taxi? 338
- 3.6** The Sugar Quota 342
- 3.7** A Tax on Gasoline 349

In §2.7, we explain that under price controls, the price of a product can be no higher than a maximum allowable ceiling price.

For a review of consumer surplus, see §4.4, where it is defined as the difference between what a consumer is willing to pay for a good and what the consumer actually pays when buying it.

is a shortage—i.e., excess demand. Of course, those consumers who can still buy the good will be better off because they will now pay less. (Presumably, this was the objective of the policy in the first place.) But if we also take into account those who cannot obtain the good, how much better off are consumers *as a whole*? Might they be worse off? And if we lump consumers and producers together, will their *total welfare* be greater or lower, and by how much? To answer questions such as these, we need a way to measure the gains and losses from government interventions and the changes in market price and quantity that such interventions cause.

Our method is to calculate the changes in *consumer and producer surplus* that result from an intervention. In Chapter 4, we saw that *consumer surplus* measures the aggregate net benefit that consumers obtain from a competitive market. In Chapter 8, we saw how *producer surplus* measures the aggregate net benefit to producers. Here we will see how consumer and producer surplus can be applied in practice.

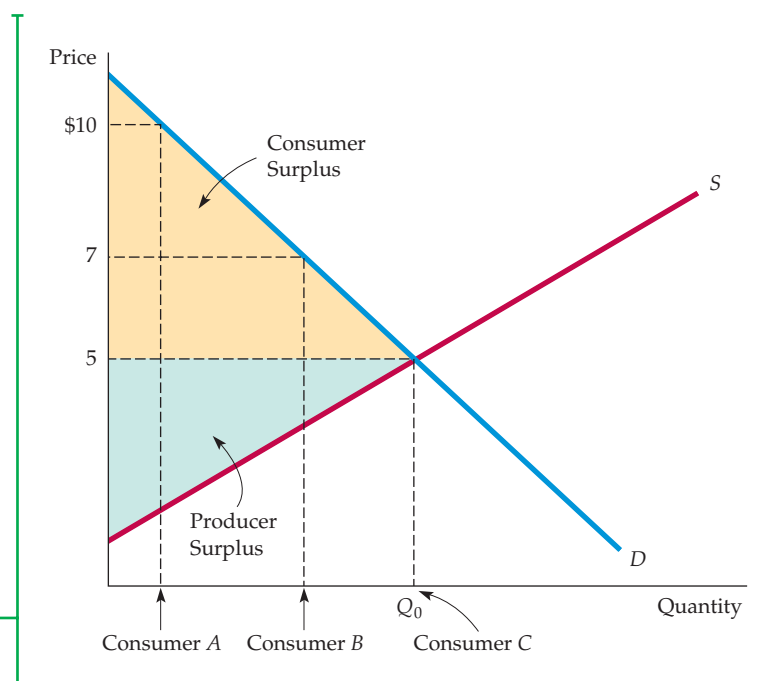
Review of Consumer and Producer Surplus

In an unregulated, competitive market, consumers and producers buy and sell at the prevailing market price. But remember, for some consumers the value of the good *exceeds* this market price; they would pay more for the good if they had to. *Consumer surplus* is the total benefit or value that consumers receive beyond what they pay for the good.

For example, suppose the market price is \$5 per unit, as in Figure 9.1. Some consumers probably value this good very highly and would pay much more than \$5 for it. Consumer A, for example, would pay up to \$10 for the good. However, because the market price is only \$5, he enjoys a net benefit of \$5—the \$10 value he places on the good, less the \$5 he must pay to obtain it. Consumer B values the good somewhat less highly. She would be willing to pay \$7, and

FIGURE 3.1
CONSUMER AND PRODUCER SURPLUS

Consumer A would pay \$10 for a good whose market price is \$5 and therefore enjoys a benefit of \$5. Consumer B enjoys a benefit of \$2, and Consumer C, who values the good at exactly the market price, enjoys no benefit. Consumer surplus, which measures the total benefit to all consumers, is the yellow-shaded area between the demand curve and the market price. Producer surplus measures the total profits of producers, plus rents to factor inputs. It is the green-shaded area between the supply curve and the market price. Together, consumer and producer surplus measure the welfare benefit of a competitive market.





thus enjoys a \$2 net benefit. Finally, Consumer C values the good at exactly the market price, \$5. He is indifferent between buying or not buying the good, and if the market price were one cent higher, he would forgo the purchase. Consumer C, therefore, obtains no net benefit.¹

For consumers in the aggregate, consumer surplus is the area between the demand curve and the market price (i.e., the yellow-shaded area in Figure 9.1). Because *consumer surplus measures the total net benefit to consumers*, we can measure the gain or loss to consumers from a government intervention by measuring the resulting change in consumer surplus.

Producer surplus is the analogous measure for producers. Some producers are producing units at a cost just equal to the market price. Other units, however, could be produced for less than the market price and would still be produced and sold even if the market price were lower. Producers, therefore, enjoy a benefit—a surplus—from selling those units. For each unit, this surplus is the difference between the market price the producer receives and the marginal cost of producing this unit.

For the market as a whole, producer surplus is the area above the supply curve up to the market price; this is *the benefit that lower-cost producers enjoy by selling at the market price*. In Figure 9.1, it is the green triangle. And because producer surplus measures the total net benefit to producers, we can measure the gain or loss to producers from a government intervention by measuring the resulting change in producer surplus.

For a review of producer surplus, see §8.6, where it is defined as the sum over all units produced of the difference between the market price of the good and the marginal cost of its production.

Application of Consumer and Producer Surplus

With consumer and producer surplus, we can evaluate the **welfare effects** of a government intervention in the market. We can determine who gains and who loses from the intervention, and by how much. To see how this is done, let's return to the example of *price controls* that we first encountered toward the end of Chapter 2. The government makes it illegal for producers to charge more than a *ceiling price* set below the market-clearing level. Recall that by decreasing production and increasing the quantity demanded, such a price ceiling creates a shortage (excess demand).

Figure 9.2 replicates Figure 2.24 (page 58), except that it also shows the changes in consumer and producer surplus that result from the government price-control policy. Let's go through these changes step by step.

1. **Change in Consumer Surplus:** Some consumers are worse off as a result of the policy, and others are better off. The ones who are worse off are those who have been rationed out of the market because of the reduction in production and sales from Q_0 to Q_1 . Other consumers, however, can still purchase the good (perhaps because they are in the right place at the right time or are willing to wait in line). These consumers are better off because they can buy the good at a lower price (P_{\max} rather than P_0).

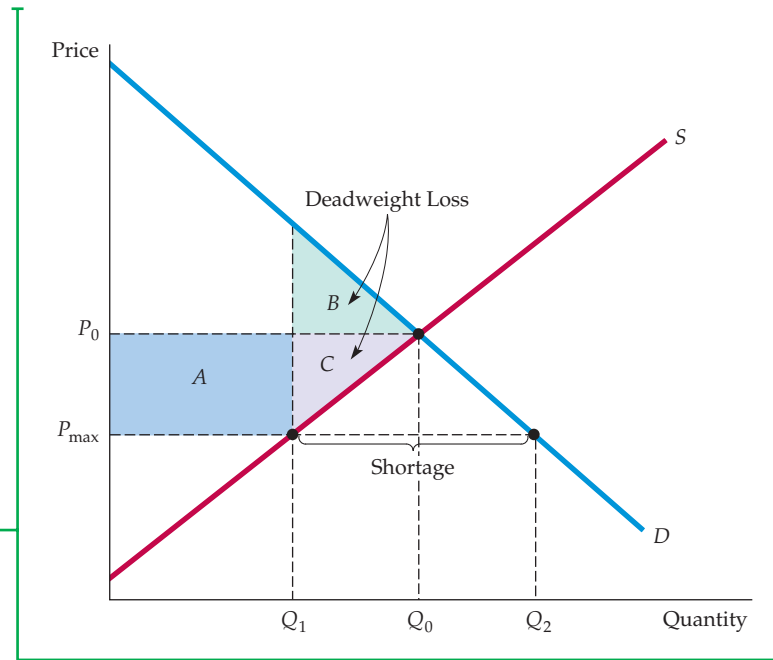
How much better off or worse off is each group? The consumers who can still buy the good enjoy an *increase* in consumer surplus, which is given by the blue-shaded rectangle A. This rectangle measures the reduction of price in each unit times the number of units consumers are able to buy at the lower price. On the other hand, those consumers who can no longer buy the good lose surplus; their *loss* is given by the green-shaded

• **welfare effects** Gains and losses to consumers and producers.

¹Of course, some consumers value the good at *less* than \$5. These consumers make up the part of the demand curve to the right of the equilibrium quantity Q_0 and will not purchase the good.

FIGURE 3.2 CHANGE IN CONSUMER AND PRODUCER SURPLUS FROM PRICE CONTROLS

The price of a good has been regulated to be no higher than P_{\max} , which is below the market-clearing price P_0 . The gain to consumers is the difference between rectangle A and triangle B . The loss to producers is the sum of rectangle A and triangle C . Triangles B and C together measure the deadweight loss from price controls.



triangle B . This triangle measures the value to consumers, less what they would have had to pay, that is lost because of the reduction in output from Q_0 to Q_1 . The net change in consumer surplus is therefore $A - B$. In Figure 9.2, because rectangle A is larger than triangle B , we know that the net change in consumer surplus is positive.

It is important to stress that we have assumed that those consumers who are able to buy the good are the ones who value it most highly. If that were not the case—e.g., if the output Q_1 were rationed randomly—the amount of lost consumer surplus would be larger than triangle B . In many cases, there is no reason to expect that those consumers who value the good most highly will be the ones who are able to buy it. As a result, the loss of consumer surplus might greatly exceed triangle B , making price controls highly inefficient.²

In addition, we have ignored the opportunity costs that arise with rationing. For example, those people who want the good might have to wait in line to obtain it. In that case, the opportunity cost of their time should be included as part of lost consumer surplus.

2. **Change in Producer Surplus:** With price controls, some producers (those with relatively lower costs) will stay in the market but will receive a lower price for their output, while other producers will leave the market. Both groups will lose producer surplus. Those who remain in the market and produce quantity Q_1 are now receiving a lower price. They have lost the producer surplus given by rectangle A . However, *total* production has also dropped. The purple-shaded triangle C measures the additional loss of producer surplus for those producers who have left the market and those

²For a nice analysis of this aspect of price controls, see David Colander, Sieuwerd Gaastra, and Casey Rothschild, "The Welfare Costs of Market Restriction," *Southern Economic Journal*, Vol. 77(1), 2011: 213–223.



who have stayed in the market but are producing less. Therefore, the total change in producer surplus is $-A - C$. Producers clearly lose as a result of price controls.

3. **Deadweight Loss:** Is the loss to producers from price controls offset by the gain to consumers? No. As Figure 9.2 shows, price controls result in a net loss of total surplus, which we call a **deadweight loss**. Recall that the change in consumer surplus is $A - B$ and that the change in producer surplus is $-A - C$. The *total* change in surplus is therefore $(A - B) + (-A - C) = -B - C$. We thus have a deadweight loss, which is given by the two triangles B and C in Figure 9.2. This deadweight loss is an inefficiency caused by price controls; the loss in producer surplus exceeds the gain in consumer surplus.

• **deadweight loss** Net loss of total (consumer plus producer) surplus.

If politicians value consumer surplus more than producer surplus, this deadweight loss from price controls may not carry much political weight. However, if the demand curve is very inelastic, price controls can result in a *net loss of consumer surplus*, as Figure 9.3 shows. In that figure, triangle B , which measures the loss to consumers who have been rationed out of the market, is larger than rectangle A , which measures the gain to consumers able to buy the good. Here, because consumers value the good highly, those who are rationed out suffer a large loss.

The demand for gasoline is very inelastic in the short run (but much more elastic in the long run). During the summer of 1979, gasoline shortages resulted from oil price controls that prevented domestic gasoline prices from increasing to rising world levels. Consumers spent hours waiting in line to buy gasoline. This was a good example of price controls making consumers—the group whom the policy was presumably intended to protect—worse off.

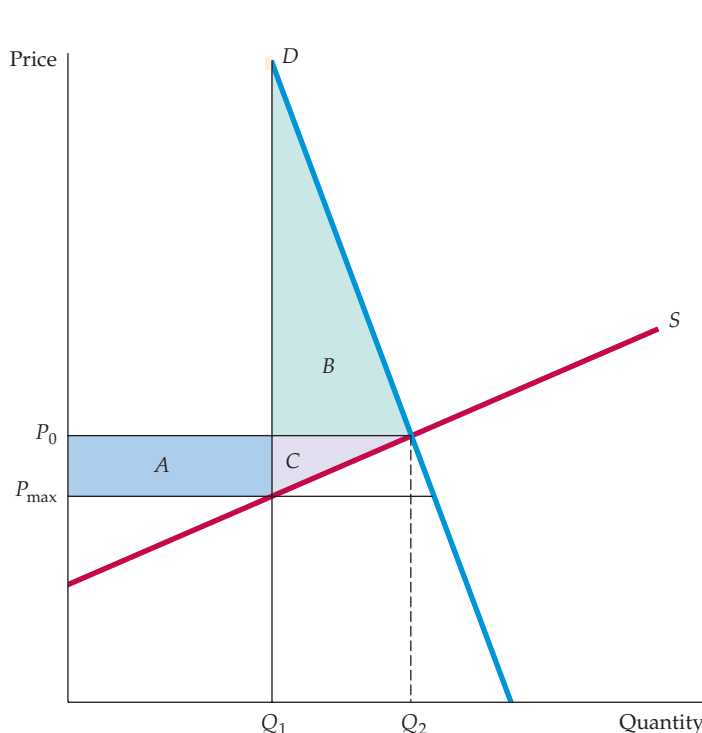


FIGURE 3.3
EFFECT OF PRICE CONTROLS WHEN DEMAND IS INELASTIC

If demand is sufficiently inelastic, triangle B can be larger than rectangle A . In this case, consumers suffer a net loss from price controls.



EXAMPLE 3.1 PRICE CONTROLS AND NATURAL GAS SHORTAGES

In Example 2.10 (page 59), we discussed the price controls that were imposed on natural gas markets during the 1970s, and we analyzed what would happen if the government were once again to regulate the wholesale price of natural gas. Specifically, we saw that, in 2007, the free-market wholesale price of natural gas was about \$6.40 per thousand cubic feet (mcf), and we calculated the quantities that would be supplied and demanded if the price were regulated to be no higher than \$3.00 per mcf. Now, equipped with the concepts of *consumer surplus*, *producer surplus*, and *deadweight loss*, we can calculate the welfare impact of this ceiling price.

Recall from Example 2.10 that we found that the supply and demand curves for natural gas could be approximated as follows:

$$\begin{aligned}\text{Supply: } Q^S &= 15.90 + 0.72P_G + 0.05P_O \\ \text{Demand: } Q^D &= 0.02 - 1.8P_G + 0.69P_O\end{aligned}$$

where Q^S and Q^D are the quantities supplied and demanded, each measured in trillion cubic feet (Tcf), P_G is the price of natural gas in dollars per thousand cubic feet (\$/mcf), and P_O is the price of oil in dollars per barrel (\$/b). As you can verify by setting Q^S equal to Q^D and using a price of oil of \$50 per barrel, the equilibrium free market price and quantity are \$6.40 per mcf and 23 Tcf, respectively. Under the hypothetical regulations, however, the maximum allowable price was \$3.00 per mcf, which implies a supply of 20.6 Tcf and a demand of 29.1 Tcf.

Figure 9.4 shows these supply and demand curves and compares the free market and regulated prices. Rectangle A and triangles B and C measure the changes in consumer and producer surplus resulting from price controls. By calculating the areas of the rectangle and triangles, we can determine the gains and losses from controls.

To do the calculations, first note that 1 Tcf is equal to 1 billion mcf. (We must put the quantities and prices in common units.) Also, by substituting the quantity 20.6 Tcf into the equation for the demand curve, we can determine that the vertical line at 20.6 Tcf intersects the demand curve at a price of \$7.73 per mcf. Then we can calculate the areas as follows:

$$\begin{aligned}A &= (20.6 \text{ billion mcf}) \times (\$3.40/\text{mcf}) = \$70.04 \text{ billion} \\ B &= (1/2) \times (2.4 \text{ billion mcf}) \times (\$1.33/\text{mcf}) = \$1.60 \text{ billion} \\ C &= (1/2) \times (2.4 \text{ billion mcf}) \times (\$3.40/\text{mcf}) = \$4.08 \text{ billion}\end{aligned}$$

(The area of a triangle is one-half the product of its altitude and its base.)

The annual change in consumer surplus that would result from these hypothetical price controls would therefore be $A - B = 70.04 - 1.60 = \$68.44$ billion. The change in producer surplus would be $-A - C = -70.04 - 4.08 = -\74.12 billion. And finally, the annual deadweight loss



would be $-B - C = -1.60 - 4.08 = -\5.68 billion. Note that most of this deadweight loss is from triangle C, i.e., the loss to those consumers who are unable to obtain natural gas as a result of the price controls.

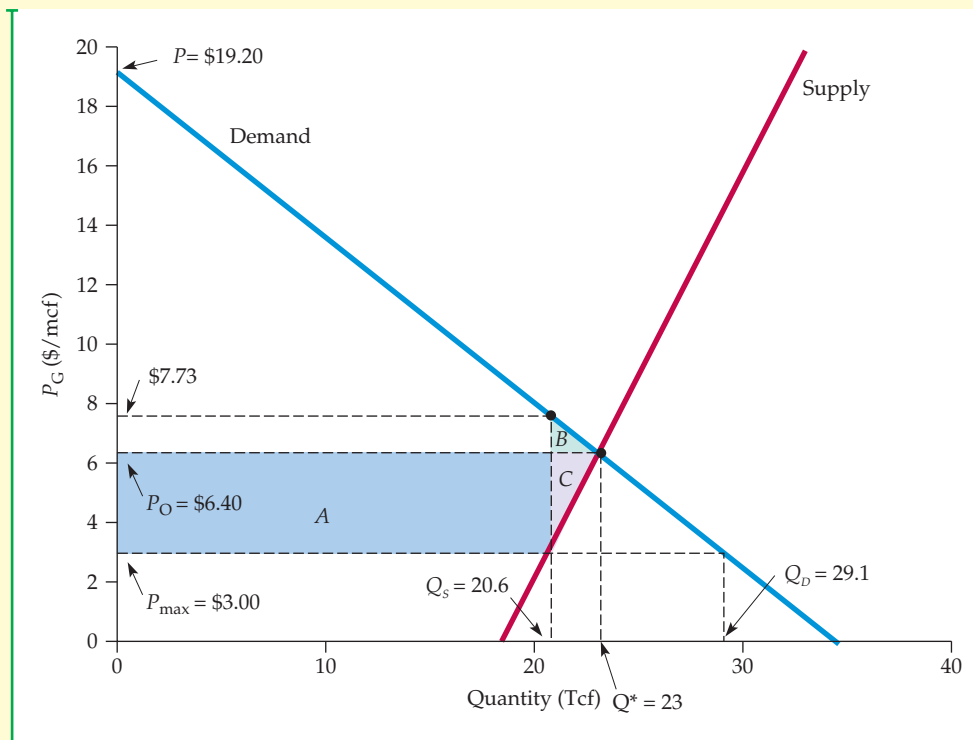


FIGURE 3.4
EFFECTS OF NATURAL GAS PRICE CONTROLS

The market-clearing price of natural gas was \$6.40 per mcf, and the (hypothetical) maximum allowable price is \$3.00. A shortage of $29.1 - 20.6 = 8.5$ Tcf results. The gain to consumers is rectangle A minus triangle B, and the loss to producers is rectangle A plus triangle C. The deadweight loss is the sum of triangles B plus C.

3.2 The Efficiency of a Competitive Market

To evaluate a market outcome, we often ask whether it achieves **economic efficiency**—the maximization of aggregate consumer and producer surplus. We just saw how price controls create a deadweight loss. The policy therefore imposes an *efficiency cost* on the economy: Taken together, producer and consumer surplus are reduced by the amount of the deadweight loss. (Of course, this does not mean that such a policy is bad; it may achieve other objectives that policymakers and the public deem important.)

• **economic efficiency**
Maximization of aggregate consumer and producer surplus.

MARKET FAILURE One might think that if the only objective is to achieve economic efficiency, a competitive market is better left alone. This is sometimes,



• **market failure** Situation in which an unregulated competitive market is inefficient because prices fail to provide proper signals to consumers and producers.

• **externality** Action taken by either a producer or a consumer which affects other producers or consumers but is not accounted for by the market price.

but not always, the case. In some situations, a **market failure** occurs: Because prices fail to provide the proper signals to consumers and producers, the unregulated competitive market is inefficient—i.e., does not maximize aggregate consumer and producer surplus. There are two important instances in which market failure can occur:

1. **Externalities:** Sometimes the actions of either consumers or producers result in costs or benefits that do not show up as part of the market price. Such costs or benefits are called **externalities** because they are “external” to the market. One example is the cost to society of environmental pollution by a producer of industrial chemicals. Without government intervention, such a producer will have no incentive to consider the social cost of pollution. We examine externalities and the proper government response to them in Chapter 18.
2. **Lack of Information:** Market failure can also occur when consumers lack information about the quality or nature of a product and so cannot make utility-maximizing purchasing decisions. Government intervention (e.g., requiring “truth in labeling”) may then be desirable. The role of information is discussed in detail in Chapter 17.

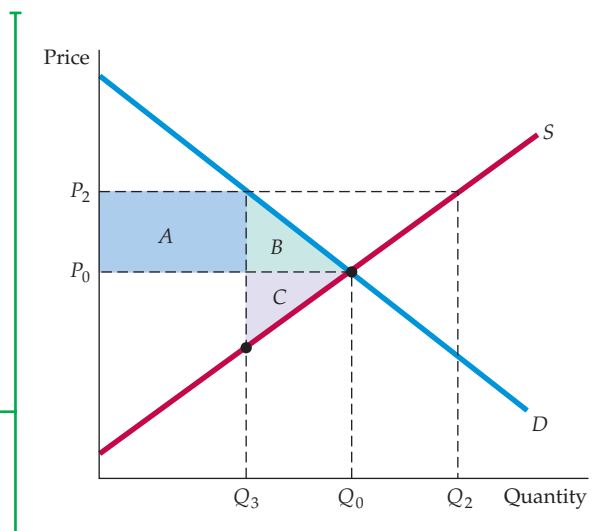
In the absence of externalities or a lack of information, an unregulated competitive market does lead to the economically efficient output level. To see this, let’s consider what happens if price is constrained to be something other than the equilibrium market-clearing price.

We have already examined the effects of a *price ceiling* (a price held below the market-clearing price). As you can see in Figure 9.2 (page 320), production falls (from Q_0 to Q_1), and there is a corresponding loss of total surplus (the deadweight-loss triangles B and C). Too little is produced, and consumers and producers in the aggregate are worse off.

Now suppose instead that the government required the price to be *above* the market-clearing price—say, P_2 instead of P_0 . As Figure 9.5 shows, although producers would like to produce more at this higher price (Q_2 instead of Q_0), consumers will now buy less (Q_3 instead of Q_0). If we assume that producers produce only what can be sold, the market output level will be Q_3 , and again, there is a net loss of total surplus. In Figure 9.5, rectangle A now represents a

FIGURE 3.5
WELFARE LOSS WHEN PRICE IS HELD ABOVE MARKET-CLEARING LEVEL

When price is regulated to be no lower than P_2 , only Q_3 will be demanded. If Q_3 is produced, the deadweight loss is given by triangles B and C. At price P_2 , producers would like to produce more than Q_3 . If they do, the deadweight loss will be even larger.





transfer from consumers to producers (who now receive a higher price), but triangles *B* and *C* again represent a deadweight loss. Because of the higher price, some consumers are no longer buying the good (a loss of consumer surplus given by triangle *B*), and some producers are no longer producing it (a loss of producer surplus given by triangle *C*).

In fact, the deadweight loss triangles *B* and *C* in Figure 9.5 give an optimistic assessment of the efficiency cost of policies that force price above market-clearing levels. Some producers, enticed by the high price P_2 , might increase their capacity and output levels, which would result in unsold output. (This happened in the airline industry when, prior to 1980, fares were regulated above market-clearing levels by the Civil Aeronautics Board.) Or to satisfy producers, the government might buy up unsold output to maintain production at Q_2 or close to it. (This is what happens in U.S. agriculture.) In both cases, the total welfare loss will exceed the areas of triangles *B* and *C*.

We will examine minimum prices, price supports, and related policies in some detail in the next few sections. Besides showing how supply–demand analysis can be used to understand and assess these policies, we will see how deviations from the competitive market equilibrium lead to efficiency costs.

EXAMPLE 3.2 THE MARKET FOR HUMAN KIDNEYS



Should people have the right to sell parts of their bodies? Congress believes the answer is no. In 1984, it passed the National Organ Transplantation Act, which prohibits the sale of organs for transplantation. Organs may only be donated.

Although the law prohibits their sale, it does not make organs valueless. Instead, it prevents those who supply organs (living persons or the families of the deceased) from reaping their economic value. It also creates a shortage of organs. Each year, about 16,000 kidneys, 44,000 corneas, and 2300 hearts are transplanted in the United

States. But there is considerable excess demand for these organs, so that many potential recipients must do without them, some of whom die as a result. For example, as of July 2011, there were about 111,500 patients on the national Organ Procurement and Transplantation Network (OPTN) waiting list. However, only 28,662 transplant surgeries were performed in the United States in 2010. Although the number of transplant surgeries has nearly doubled since 1990, the number of patients waiting for organs has increased to nearly five times its level in 1990.³

To understand the effects of this law, let's consider the supply and demand for kidneys. First the supply curve. Even at a price of zero (the effective price under the law), donors supply about 16,000 kidneys per

³Source: Organ Procurement and Transplantation Network, <http://www.optn.transplant.hrsa.gov>.



year. But many other people who need kidney transplants cannot obtain them because of a lack of donors. It has been estimated that 8000 more kidneys would be supplied if the price were \$20,000. We can fit a linear supply curve to this data—i.e., a supply curve of the form $Q = a + bP$. When $P = 0$, $Q = 16,000$, so $a = 16,000$. If $P = \$20,000$, $Q = 24,000$, so $b = (24,000 - 16,000)/20,000 = 0.4$. Thus the supply curve is

$$\text{Supply: } Q^S = 16,000 + 0.4P$$

Note that at a price of \$20,000, the elasticity of supply is 0.33.

It is expected that at a price of \$20,000, the number of kidneys demanded would be 24,000 per year. Like supply, demand is relatively price inelastic; a reasonable estimate for the price elasticity of demand at the \$20,000 price is -0.33 . This implies the following linear demand curve:

$$\text{Demand: } Q^D = 32,000 - 0.4P$$

These supply and demand curves are plotted in Figure 9.6, which shows the market-clearing price and quantity of \$20,000 and 24,000, respectively.

In §2.6, we explain how to fit linear demand and supply curves from information about the equilibrium price and quantity and the price elasticities of demand and supply.

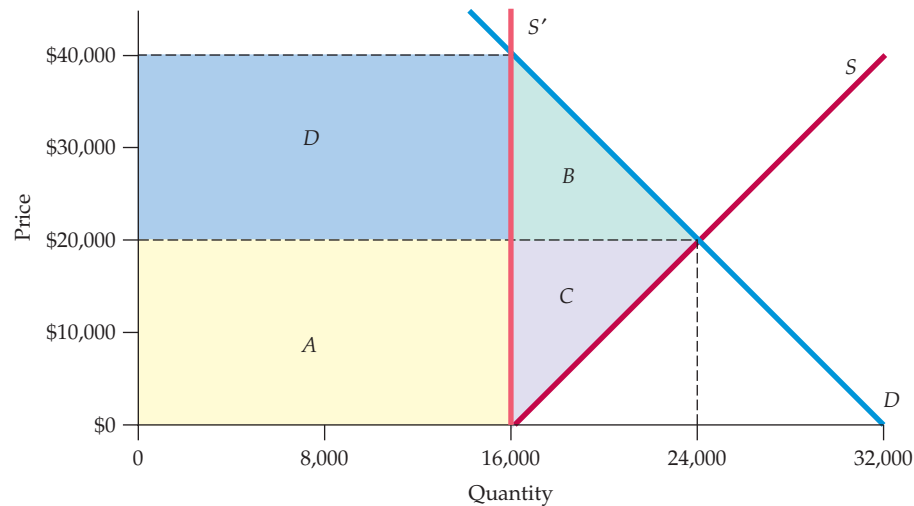


FIGURE 3.6
THE MARKET FOR KIDNEYS AND THE EFFECT OF THE NATIONAL ORGAN TRANSPLANTATION ACT

The market-clearing price is \$20,000; at this price, about 24,000 kidneys per year would be supplied. The law effectively makes the price zero. About 16,000 kidneys per year are still donated; this constrained supply is shown as S' . The loss to suppliers is given by rectangle A and triangle C. If consumers received kidneys at no cost, their gain would be given by rectangle A less triangle B. In practice, kidneys are often rationed on the basis of willingness to pay, and many recipients pay most or all of the \$40,000 price that clears the market when supply is constrained. Rectangles A and D measure the total value of kidneys when supply is constrained.



Because the sale of kidneys is prohibited, supply is limited to 16,000 (the number of kidneys that people donate). This constrained supply is shown as the vertical line S' . How does this affect the welfare of kidney suppliers and recipients?

First consider suppliers. Those who provide kidneys fail to receive the \$20,000 that each kidney is worth—a loss of surplus represented by rectangle A and equal to $(16,000)(\$20,000) = \320 million. Moreover, some people who would supply kidneys if they were paid do not. These people lose an amount of surplus represented by triangle C, which is equal to $(1/2)(8000)(\$20,000) = \80 million. Therefore, the total loss to suppliers is \$400 million.

What about recipients? Presumably the law intended to treat the kidney as a gift to the recipient. In this case, those recipients who obtain kidneys *gain* rectangle A (\$320 million) because they (or their insurance companies) do not have to pay the \$20,000 price. Those who cannot obtain kidneys lose surplus of an amount given by triangle B and equal to \$80 million. This implies a net increase in the surplus of recipients of $\$320 \text{ million} - \$80 \text{ million} = \$240 \text{ million}$. It also implies a deadweight loss equal to the areas of triangles B and C (i.e., \$160 million).

These estimates of the welfare effects of the policy may need adjustment for two reasons. First, kidneys will not necessarily be allocated to those who value them most highly. If the limited supply of kidneys is partly allocated to people with valuations below \$40,000, the true deadweight loss will be higher than our estimate. Second, with excess demand, there is no way to ensure that recipients will receive their kidneys as gifts. In practice, kidneys are often rationed on the basis of willingness to pay, and many recipients end up paying all or most of the \$40,000 price that is needed to clear the market when supply is constrained to 16,000. A good part of the value of the kidneys—rectangles A and D in the figure—is then captured by hospitals and middlemen. As a result, the law reduces the surplus of recipients as well as of suppliers.⁴

There are, of course, arguments in favor of prohibiting the sale of organs.⁵ One argument stems from the problem of imperfect information; if people receive payment for organs, they may hide adverse information about their health histories. This argument is probably most applicable to the sale of blood, where there is a possibility of transmitting hepatitis, AIDS, or other viruses. But even in such cases, screening (at a cost that would be included in the market price) may be more efficient than prohibiting sales. This issue has been central to the debate in the United States over blood policy.

A second argument holds that it is simply unfair to allocate a basic necessity of life on the basis of ability to pay. This argument transcends economics.

⁴For further analyses of these efficiency costs, see Dwane L. Barney and R. Larry Reynolds, "An Economic Analysis of Transplant Organs," *Atlantic Economic Journal* 17 (September 1989): 12–20; David L. Kaserman and A. H. Barnett, "An Economic Analysis of Transplant Organs: A Comment and Extension," *Atlantic Economic Journal* 19 (June 1991): 57–64; and A. Frank Adams III, A. H. Barnett, and David L. Kaserman, "Markets for Organs: The Question of Supply," *Contemporary Economic Policy* 17 (April 1999): 147–55. Kidney exchange is also complicated by the need to match blood type; for a recent analysis, see Alvin E. Roth, Tayfun Sönmez, and M. Utku Ünver, "Efficient Kidney Exchange: Coincidence of Wants in Markets with Compatibility-Based Preferences," *American Economic Review* 97 (June 2007).

⁵For discussions of the strengths and weaknesses of these arguments, see Susan Rose-Ackerman, "Inalienability and the Theory of Property Rights," *Columbia Law Review* 85 (June 1985): 931–69, and Roger D. Blair and David L. Kaserman, "The Economics and Ethics of Alternative Cadaveric Organ Procurement Policies," *Yale Journal on Regulation* 8 (Summer 1991): 403–52.



However, two points should be kept in mind. First, when the price of a good that has a significant opportunity cost is forced to zero, there is bound to be reduced supply and excess demand. Second, it is not clear why live organs should be treated differently from close substitutes; artificial limbs, joints, and heart valves, for example, are sold even though real kidneys are not.

Many complex ethical and economic issues are involved in the sale of organs. These issues are important, and this example is not intended to sweep them away. Economics, the dismal science, simply shows us that human organs have economic value that cannot be ignored, and that prohibiting their sale imposes a cost on society that must be weighed against the benefits.

3.3 Minimum Prices

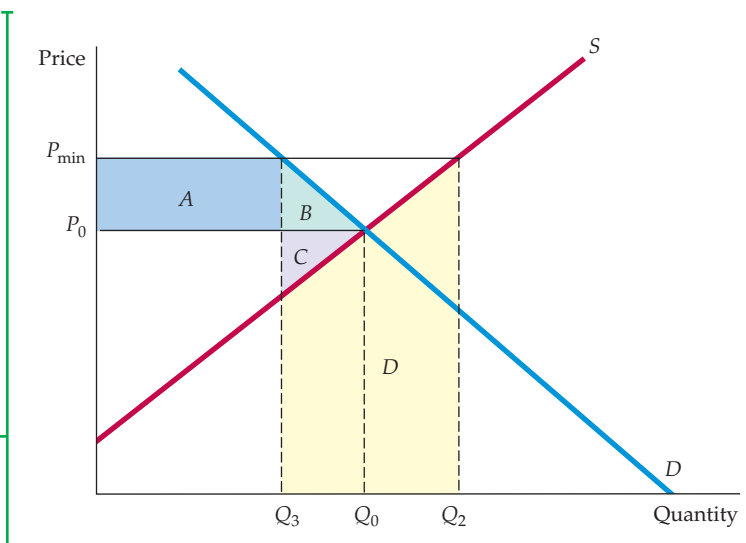
As we have seen, government policy sometimes seeks to *raise* prices above market-clearing levels, rather than lower them. Examples include the former regulation of the airlines by the Civil Aeronautics Board, the minimum wage law, and a variety of agricultural policies. (Most import quotas and tariffs also have this intent, as we will see in Section 9.5.) One way to raise prices above market-clearing levels is by direct regulation—simply make it illegal to charge a price lower than a specific minimum level.

Look again at Figure 9.5 (page 324). If producers correctly anticipate that they can sell only the lower quantity Q_3 , the net welfare loss will be given by triangles B and C . But as we explained, producers might not limit their output to Q_3 . What happens if producers think they can sell all they want at the higher price and produce accordingly? That situation is illustrated in Figure 9.7, where P_{\min} denotes a minimum price set by the government. The quantity supplied is now Q_2 and the quantity demanded is Q_3 , the difference representing excess, unsold supply. Now let's determine the resulting changes in consumer and producer surplus.

Those consumers who still purchase the good must now pay a higher price and so suffer a loss of surplus, which is given by rectangle A in Figure 9.7. Some

FIGURE 3.7
PRICE MINIMUM

Price is regulated to be no lower than P_{\min} . Producers would like to supply Q_2 , but consumers will buy only Q_3 . If producers indeed produce Q_2 , the amount $Q_2 - Q_3$ will go unsold and the change in producer surplus will be $A - C - D$. In this case, producers as a group may be worse off.





consumers have also dropped out of the market because of the higher price, with a corresponding loss of surplus given by triangle *B*. The total change in consumer surplus is therefore

$$\Delta CS = -A - B$$

Consumers clearly are worse off as a result of this policy.

What about producers? They receive a higher price for the units they sell, which results in an increase of surplus, given by rectangle *A*. (Rectangle *A* represents a transfer of money from consumers to producers.) But the drop in sales from Q_0 to Q_3 results in a loss of surplus, which is given by triangle *C*. Finally, consider the cost to producers of expanding production from Q_0 to Q_2 . Because they sell only Q_3 , there is no revenue to cover the cost of producing $Q_2 - Q_3$. How can we measure this cost? Remember that the supply curve is the aggregate marginal cost curve for the industry. The supply curve therefore gives us the additional cost of producing each incremental unit. Thus the area under the supply curve from Q_3 to Q_2 is the cost of producing the quantity $Q_2 - Q_3$. This cost is represented by the shaded trapezoid *D*. So unless producers respond to unsold output by cutting production, the total change in producer surplus is

$$\Delta PS = A - C - D$$

Given that trapezoid *D* can be large, a minimum price can even result in a net loss of surplus to producers alone! As a result, this form of government intervention can reduce producers' profits because of the cost of excess production.

Another example of a government-imposed price minimum is a minimum wage law. The effect of this policy is illustrated in Figure 9.8, which shows the supply and demand for labor. The wage is set at w_{\min} , a level higher than the market-clearing wage w_0 . As a result, those workers who can find jobs obtain a higher wage. However, some people who want to work will be unable to. The policy results in unemployment, which in the figure is $L_2 - L_1$. We will examine the minimum wage in more detail in Chapter 14.

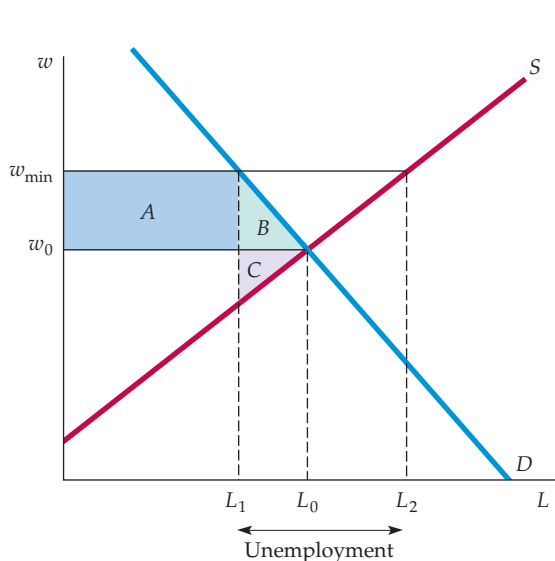


FIGURE 9.8
THE MINIMUM WAGE

Although the market-clearing wage is w_0 , firms are not allowed to pay less than w_{\min} . This results in unemployment of an amount $L_2 - L_1$ and a deadweight loss given by triangles *B* and *C*.



EXAMPLE 3.3 AIRLINE REGULATION

Before 1980, the airline industry in the United States looked very different than it does today. Fares and routes were tightly regulated by the Civil Aeronautics Board (CAB). The CAB set most fares well above what would have prevailed in a free market. It also restricted entry, so that many routes were served by only one or two airlines. By the late 1970s, however, the CAB liberalized fare regulation and allowed airlines to serve any routes they wished. By 1981, the industry had been completely deregulated, and the CAB itself was dissolved in 1982. Since that time, many new airlines have begun service, others have gone out of business, and price competition has become much more intense.

Many airline executives feared that deregulation would lead to chaos in the industry, with competitive pressure causing sharply reduced profits and even bankruptcies. After all, the original rationale for CAB

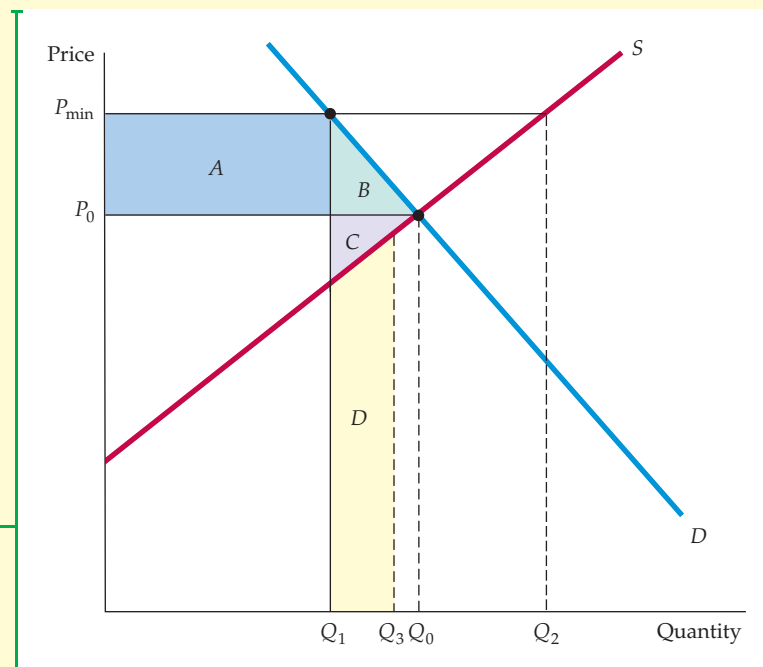


regulation was to provide “stability” in an industry that was considered vital to the U.S. economy. And one might think that as long as price was held above its market-clearing level, profits would be higher than they would be in a free market.

Deregulation did lead to major changes in the industry. Some airlines merged or went out of business as new ones entered. Although prices fell considerably (to the benefit of consumers), profits overall did not fall much because the CAB’s minimum prices had caused inefficiencies and artificially high costs. The effect of minimum prices is illustrated in Figure 9.9, where P_0 and Q_0 are the market-clearing price and quantity, P_{\min} is the minimum price, and Q_1 is the amount demanded at this higher price. The problem was that at price P_{\min} , airlines wanted to supply a quantity Q_2 , much larger than Q_1 . Although they did not expand output to Q_2 , they did expand it well beyond Q_1 —to Q_3 in the figure—hoping to

FIGURE 3.9
EFFECT OF AIRLINE REGULATION BY THE CIVIL AERONAUTICS BOARD

At price P_{\min} , airlines would like to supply Q_2 , well above the quantity Q_1 that consumers will buy. Here they supply Q_3 . Trapezoid D is the cost of unsold output. Airline profits may have been lower as a result of regulation because triangle C and trapezoid D can together exceed rectangle A . In addition, consumers lose $A + B$.





sell this quantity at the expense of competitors. As a result, load factors (the percentage of seats filled) were relatively low, and so were profits. (Trapezoid *D* measures the cost of unsold output.)

Table 9.1 gives some key numbers that illustrate the evolution of the airline industry.⁶ The number of carriers increased dramatically after deregulation, as did passenger load factors (the percentage of seats with passengers). The passenger-mile rate (the revenue per passenger-mile flown) fell sharply in real (inflation-adjusted) terms from 1980 to 1990, and then continued to drop through 2010. This decline was the result of increased competition and reductions in fares, and made air travel affordable to many more consumers.

And what about costs? The real cost index indicates that even after adjusting for inflation, costs increased by about 45 percent between 1975 and 1980, and then fell considerably over the next 20 years. Changes in cost, however, are driven to a great extent by changes in the cost of fuel, which is driven in turn by changes in the price of oil. (For most airlines, fuel accounts for close to 30 percent of total operating costs.) As Table 9.1 shows, the real cost of fuel has fluctuated dramatically, and this had nothing to do with deregulation.

Because airlines have no control over oil prices, it is more informative to examine a “corrected” real cost index which removes the effects of changing fuel costs. Real fuel costs increased considerably from 1975 to 1980, which accounts for much of the increase in the real cost index. Real fuel costs nearly tripled from 2000 to 2010 (because of sharp increases in the price of oil); had fuel costs remained level, the real cost index would have *declined* (from 85 to 76) rather than increasing sharply (from 89 to 148).

What, then, did airline deregulation do for consumers and producers? As new airlines entered the industry and fares went down, consumers benefited. This fact is borne out by the increase in consumer surplus given by rectangle *A* and triangle *B* in Figure 9.9. (The actual benefit to consumers was somewhat smaller because *quality* declined as planes became more crowded and delays and cancellations multiplied.) As for the airlines, they had to learn to live in a more competitive—and therefore more turbulent—environment, and some firms did not survive. But overall, airlines became so much more efficient that producer surplus may have increased. The total welfare gain from deregulation was positive and quite large.⁷

TABLE 3.1 AIRLINE INDUSTRY DATA

	1975	1980	1990	2000	2010
Number of U.S. Carriers	36	63	70	94	63
Passenger Load Factor (%)	54.0	58.0	62.4	72.1	82.1
Passenger-Mile Rate (constant 1995 dollars)	0.218	0.210	0.149	0.118	0.094
Real Cost Index (1995 = 100)	101	145	119	89	148
Real Fuel Cost Index (1995 = 100)	249	300	163	125	342
Real Cost Index w/o Fuel Cost Increases (1995 = 100)	71	87	104	85	76

⁶Department of Commerce, Air Transport Association.

⁷Studies of the effects of deregulation include John M. Trapani and C. Vincent Olson, “An Analysis of the Impact of Open Entry on Price and the Quality of Service in the Airline Industry,” *Review of Economics and Statistics* 64 (February 1982): 118–38; David R. Graham, Daniel P. Kaplan, and David S. Sibley, “Efficiency and Competition in the Airline Industry,” *Bell Journal of Economics* (Spring 1983): 118–38; S. Morrison and Clifford Winston, *The Economic Effects of Airline Deregulation* (Washington: Brookings Institution, 1986); and Nancy L. Rose, “Profitability and Product Quality: Economic Determinants of Airline Safety Performance,” *Journal of Political Economy* 98 (October 1990): 944–64.



• **price support** Price set by government above free-market level and maintained by governmental purchases of excess supply.

3.4 Price Supports and Production Quotas

Besides imposing a minimum price, the government can increase the price of a good in other ways. Much of American agricultural policy is based on a system of **price supports**, whereby the government sets the market price of a good above the free-market level and buys up whatever output is needed to maintain that price. The government can also increase prices by *restricting production*, either directly or through incentives to producers. In this section, we show how these policies work and examine their impact on consumers, producers, and the federal budget.

Price Supports

In the United States, price supports aim to increase the prices of dairy products, tobacco, corn, peanuts, and so on, so that the producers of those goods can receive higher incomes. Under a price support program, the government sets a support price P_s and then buys up whatever output is needed to keep the market price at this level. Figure 9.10 illustrates this. Let's examine the resulting gains and losses to consumers, producers, and the government.

CONSUMERS At price P_s , the quantity that consumers demand falls to Q_1 , but the quantity supplied increases to Q_2 . To maintain this price and avoid having inventories pile up in producer warehouses, the government must buy the quantity $Q_g = Q_2 - Q_1$. In effect, because the government adds its demand Q_g to the demand of consumers, producers can sell all they want at price P_s .

Because those consumers who purchase the good must pay the higher price P_s instead of P_o , they suffer a loss of consumer surplus given by rectangle A . Because of the higher price, other consumers no longer buy the good or buy less of it, and their loss of surplus is given by triangle B . So, as with the minimum price that we examined above, consumers lose, in this case by an amount

$$\Delta CS = -A - B$$

PRODUCERS On the other hand, producers gain (which is why such a policy is implemented). Producers are now selling a larger quantity Q_2 instead of Q_o , and at a higher price P_s . Observe from Figure 9.10 that producer surplus increases by the amount

$$\Delta PS = A + B + D$$

THE GOVERNMENT But there is also a cost to the government (which must be paid for by taxes, and so is ultimately a cost to consumers). That cost is $(Q_2 - Q_1)P_s$, which is what the government must pay for the output it purchases. In Figure 9.10, this amount is represented by the large speckled rectangle. This cost may be reduced if the government can “dump” some of its purchases—i.e., sell them abroad at a low price. Doing so, however, hurts the ability of domestic producers to sell in foreign markets, and it is domestic producers that the government is trying to please in the first place.

What is the total welfare cost of this policy? To find out, we add the change in consumer surplus to the change in producer surplus and then subtract the cost to the government. Thus the total change in welfare is

$$\Delta CS + \Delta PS - \text{Cost to Govt.} = D - (Q_2 - Q_1)P_s$$

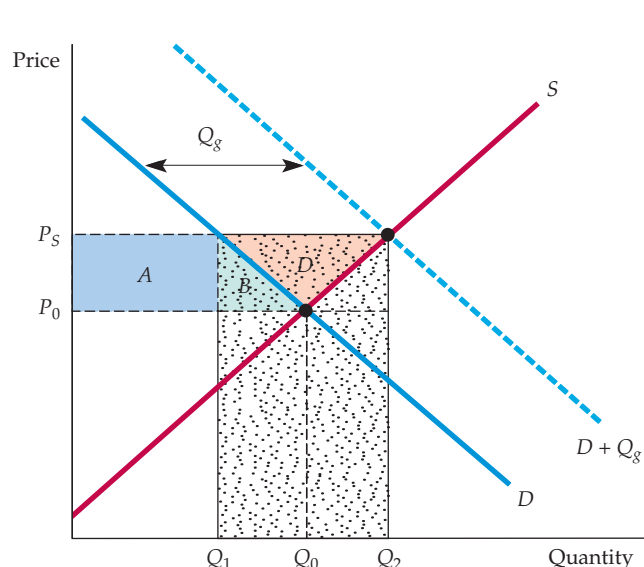


FIGURE 3.10
PRICE SUPPORTS

To maintain a price P_s above the market-clearing price P_0 , the government buys a quantity Q_g . The gain to producers is $A + B + D$. The loss to consumers is $A + B$. The cost to the government is the speckled rectangle, the area of which is $P_s(Q_2 - Q_1)$.

In terms of Figure 9.10, society as a whole is worse off by an amount given by the large speckled rectangle, less triangle D .

As we will see in Example 9.4, this welfare loss can be very large. But the most unfortunate part of this policy is the fact that there is a much more efficient way to help farmers. If the objective is to give farmers an additional income equal to $A + B + D$, it is far less costly to society to give them this money directly rather than via price supports. Because price supports are costing consumers $A + B$ anyway, by paying farmers directly, society saves the large speckled rectangle, less triangle D . So why doesn't the government simply give farmers money? Perhaps because price supports are a less obvious giveaway and, therefore, politically more attractive.⁸

Production Quotas

Besides entering the market and buying up output—thereby increasing total demand—the government can also cause the price of a good to rise by *reducing supply*. It can do this by decree—that is, by simply setting quotas on how much each firm can produce. With appropriate quotas, the price can then be forced up to any arbitrary level.

As we will see in Example 9.5, this is how many city governments maintain high taxi fares. They limit total supply by requiring each taxicab to have a medallion, and then limit the total number of medallions. Another example is the control of liquor licenses by state governments. By requiring any bar or restaurant that serves alcohol to have a liquor license and then limiting the number of licenses, entry by new restaurateurs is limited, which allows those who have licenses to earn higher prices and profit margins.

⁸In practice, price supports for many agricultural commodities are effected through loans. The loan rate is in effect a price floor. If during the loan period market prices are not sufficiently high, farmers can forfeit their grain to the government (specifically to the Commodity Credit Corporation) as *full payment for the loan*. Farmers have the incentive to do this unless the market price rises above the support price.



The welfare effects of production quotas are shown in Figure 9.11. The government restricts the quantity supplied to Q_1 , rather than the market-clearing level Q_0 . Thus the supply curve becomes the vertical line S' at Q_1 . Consumer surplus is reduced by rectangle A (those consumers who buy the good pay a higher price) plus triangle B (at this higher price, some consumers no longer purchase the good). Producers gain rectangle A (by selling at a higher price) but lose triangle C (because they now produce and sell Q_1 rather than Q_0). Once again, there is a deadweight loss, given by triangles B and C .

INCENTIVE PROGRAMS In U.S. agricultural policy, output is reduced by incentives rather than by outright quotas. *Acreage limitation programs* give farmers financial incentives to leave some of their acreage idle. Figure 9.11 also shows the welfare effects of reducing supply in this way. Note that because farmers agree to limit planted acreage, the supply curve again becomes completely inelastic at the quantity Q_1 , and the market price is increased from P_0 to P_s .

As with direct production quotas, the change in consumer surplus is

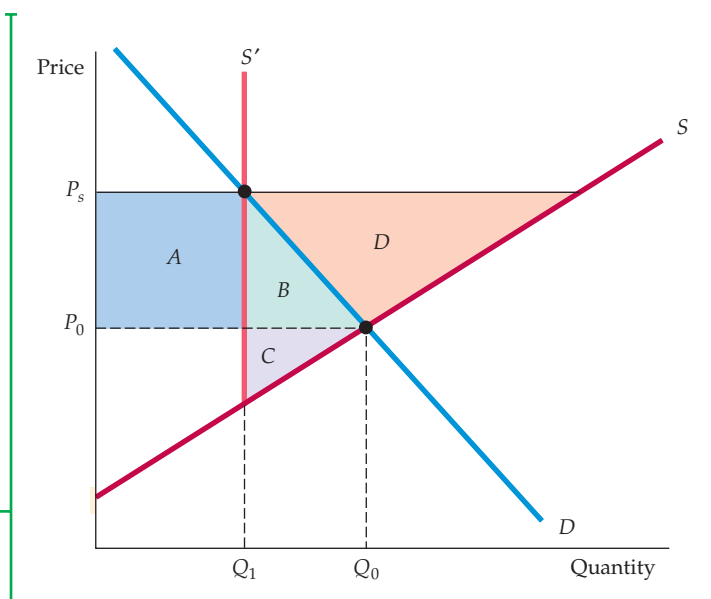
$$\Delta CS = -A - B$$

Farmers now receive a higher price for the production Q_1 , which corresponds to a gain in surplus of rectangle A . But because production is reduced from Q_0 to Q_1 , there is a loss of producer surplus corresponding to triangle C . Finally, farmers receive money from the government as an incentive to reduce production. Thus the total change in producer surplus is now

$$\Delta PS = A - C + \text{Payments for not producing}$$

FIGURE 3.11 SUPPLY RESTRICTIONS

To maintain a price P_s above the market-clearing price P_0 , the government can restrict supply to Q_1 , either by imposing production quotas (as with taxicab medallions) or by giving producers a financial incentive to reduce output (as with acreage limitations in agriculture). For an incentive to work, it must be at least as large as $B + C + D$, which would be the additional profit earned by planting, given the higher price P_s . The cost to the government is therefore at least $B + C + D$.





The cost to the government is a payment sufficient to give farmers an incentive to reduce output to Q_1 . That incentive must be at least as large as $B + C + D$ because that area represents the additional profit that could be made by planting, *given the higher price P_s* . (Remember that the higher price P_s gives farmers an incentive to produce *more* even though the government is trying to get them to produce *less*.) Thus the cost to the government is at least $B + C + D$, and the total change in producer surplus is

$$\Delta PS = A - C + B + C + D = A + B + D$$

This is the same change in producer surplus as with price supports maintained by government purchases of output. (Refer to Figure 9.10.) Farmers, then, should be indifferent between the two policies because they end up gaining the same amount of money from each. Likewise, consumers lose the same amount of money.

Which policy costs the government more? The answer depends on whether the sum of triangles $B + C + D$ in Figure 9.11 is larger or smaller than $(Q_2 - Q_1)P_s$ (the large speckled rectangle) in Figure 9.10. Usually it will be smaller, so that an acreage-limitation program costs the government (and society) less than price supports maintained by government purchases.

Still, even an acreage-limitation program is more costly to society than simply handing the farmers money. The total change in welfare ($\Delta CS + \Delta PS - \text{Cost to Govt.}$) under the acreage-limitation program is

$$\Delta \text{Welfare} = -A - B + A + B + D - B - C - D = -B - C$$

Society would clearly be better off in efficiency terms if the government simply gave the farmers $A + B + D$, leaving price and output alone. Farmers would then gain $A + B + D$ and the government would lose $A + B + D$, for a total welfare change of zero, instead of a loss of $B + C$. However, economic efficiency is not always the objective of government policy.

EXAMPLE 3.4 SUPPORTING THE PRICE OF WHEAT

In Examples 2.5 (page 37) and 4.3 (page 128), we began to examine the market for wheat in the United States. Using linear demand and supply curves, we found that the market-clearing price of wheat was about \$3.46 in 1981. The price fell to about \$2.78 by 2002 because of a drop in export demand. In fact, government programs kept the actual price of wheat higher and provided direct subsidies to farmers. How did these programs work, how much did they end up costing consumers, and how much did they add to the federal deficit?





First, let's examine the market in 1981. In that year, although there were no effective limitations on the production of wheat, the price was increased to \$3.70 by government purchases. How much would the government have had to buy to get the price from \$3.46 to \$3.70? To answer this question, first write the equations for supply and for total private (domestic plus export) demand:

$$1981 \text{ Supply: } Q_s = 1800 + 240P$$

$$1981 \text{ Demand: } Q_D = 3550 - 266P$$

By equating supply and demand, you can check that the market-clearing price is \$3.46, and that the quantity produced is 2630 million bushels. Figure 9.12 illustrates this.

To increase the price to \$3.70, the government must buy a quantity of wheat Q_g . Total demand (private plus government) will then be

$$1981 \text{ Total demand: } Q_{DT} = 3550 - 266P + Q_g$$

Now equate supply with this total demand:

$$1800 + 240P = 3550 - 266P + Q_g$$

or

$$Q_g = 506P - 1750$$

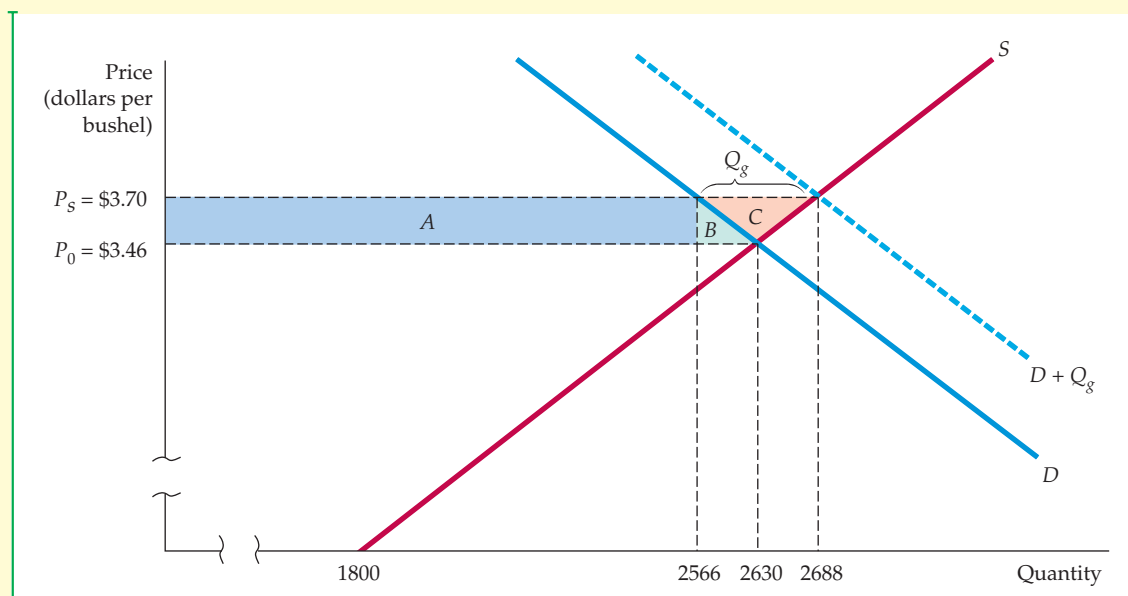


FIGURE 3.12
THE WHEAT MARKET IN 1981

By buying 122 million bushels of wheat, the government increased the market-clearing price from \$3.46 per bushel to \$3.70.



This equation can be used to determine the required quantity of government wheat purchases Q_g as a function of the desired support price P . To achieve a price of \$3.70, the government must buy

$$Q_g = (506)(3.70) - 1750 = 122 \text{ million bushels}$$

Note in Figure 9.12 that these 122 million bushels are the difference between the quantity supplied at the \$3.70 price (2688 million bushels) and the quantity of private demand (2566 million bushels). The figure also shows the gains and losses to consumers and producers. Recall that consumers lose rectangle *A* and triangle *B*. You can verify that rectangle *A* is $(3.70 - 3.46)(2566) = \$616$ million, and triangle *B* is $(1/2)(3.70 - 3.46)(2630 - 2566) = \8 million, so that the total cost to consumers is \$624 million.

The cost to the government is the \$3.70 it pays for the wheat times the 122 million bushels it buys, or \$451.4 million. The total cost of the program is then \$624 million + \$451.4 million = \$1075 million. Compare this with the gain to producers, which is rectangle *A* plus triangles *B* and *C*. You can verify that this gain is \$638 million.

Price supports for wheat were expensive in 1981. To increase the surplus of farmers by \$638 million, consumers and taxpayers had to pay \$1076 million. In fact, taxpayers paid even more than that. Wheat producers were also given subsidies of about 30 cents per bushel, which adds up to another \$806 million.

In 1996, the U.S. Congress passed a new farm bill, nicknamed the "Freedom to Farm" law. It was designed to reduce the role of government and to make agriculture more market oriented. The law eliminated production quotas (for wheat, corn, rice, and other products) and gradually reduced government purchases and subsidies through 2003. However, the law did not completely deregulate U.S. agriculture. For example, price support programs for peanuts and sugar remained in place. Furthermore, pre-1996 price supports and production quotas would be reinstated unless Congress renewed the law in 2003. (Congress did not renew it—more on this below.) Even under the 1996 law, agricultural subsidies remained substantial.

In Example 2.5, we saw that the market-clearing price of wheat in 2007 had increased to about \$6.00 per bushel. The supply and demand curves in 2007 were as follows:

$$\text{Demand: } Q_D = 2900 - 125P$$

$$\text{Supply: } Q_S = 1460 + 115P$$

You can check to see that the market-clearing quantity is 2150 million bushels.

Congress did not renew the 1996 Freedom to Farm Act. Instead, in 2002, Congress and the Bush administration essentially reversed the effects of the 1996 bill through passage of the Farm Security and Rural Investment Act, which reinstates subsidies for most crops, in particular grain and cotton.⁹

⁹See Mike Allen, "Bush Signs Bill Providing Big Farm Subsidy Increases," *The Washington Post*, May 14, 2002; see David E. Sanger, "Reversing Course, Bush Signs Bill Raising Farm Subsidies," *The New York Times*, May 14, 2002.



Although the bill does not explicitly restore price supports, it calls for the government to issue “fixed direct payments” to producers based on a fixed payment rate and the base acreage for a particular crop. Using U.S. wheat acreage and production levels in 2001, we can calculate that this bill cost taxpayers nearly \$1.1 billion in annual payments to wheat producers alone.¹⁰ The 2002 farm bill was projected to cost taxpayers \$190 billion over 10 years.

Congress revisited agricultural subsidies in 2007. For most crops, previous subsidy rates were either maintained or increased, thus making the burden on U.S. taxpayers even higher. In fact, the Food, Conservation, and Energy Act of 2008 raised subsidy rates on most crops through 2012, at a projected cost of \$284 billion over five years. Recently, however, the pendulum has swung back toward eliminating subsidies, and new cuts were approved as part of the deal to resolve the 2011 budget crisis.

EXAMPLE 3.5 WHY CAN'T I FIND A TAXI?

Ever try to catch a cab in New York? Good luck! If it's raining or it's a peak commuting time, you can wait an hour before successfully hailing a cab. Why? Why aren't there more taxis in New York?

The reason is simple. The city of New York limits the number of taxis by requiring each taxi to have a *medallion* (essentially a permit), and then limiting the number of medallions. In 2011 there were 13,150 medallions in New York—roughly the same number as in 1937, a time when it was much easier to find a taxi. But since 1937 the city has grown and the demand for taxi rides has increased greatly, so that now the limit of 13,150 medallions is a constraint that can make life difficult for New Yorkers. But that just raises another question. Why would a city do something that makes life difficult for its citizens? Why not just issue more medallions?

Again, the reason is simple. Doing so would incur the wrath of the current owners of medallions—mostly large taxi companies that lease the medal-

lions and taxis to drivers, and have considerable political and lobbying power. Medallions can be bought and sold by the companies that own them. In 1937, there were plenty of medallions to go around, so they had little value. By 1947, the value of a medallion had increased to \$2,500, by 1980 to \$55,000, and by 2011 to \$880,000. That's right—because New York City won't issue more medallions, the value of a taxi medallion is approaching \$1 million! But of course that value would drop sharply if the city starting issuing more medallions. So the New York taxi companies that collectively own the 13,150 available medallions have done everything possible to prevent the city from issuing any more—and have succeeded in their efforts.

The situation is illustrated in Figure 9.13. The demand curve D and supply curve S are based on elasticities taken from statistical studies of taxicab markets in New York and other cities.¹¹ If the

¹⁰Estimated 2001 Wheat direct payments = (payment rate)*(payment yield)*(base acres)* 0.85 = $(\$0.52)*(40.2)*(59,617,000)*0.85 = \1.06 billion.

¹¹Elasticities are taken from Bruce Schaller, “Elasticities for Taxicab Fares and Service Availability,” *Transportation* 26 (1999): 283–297. Information about New York's taxi regulations and medallion prices can be found at New York City's Taxi and Limousine Commission's website: <http://www.nyc.gov/tlc>, and at <http://www.schallerconsult.com/taxi/>.

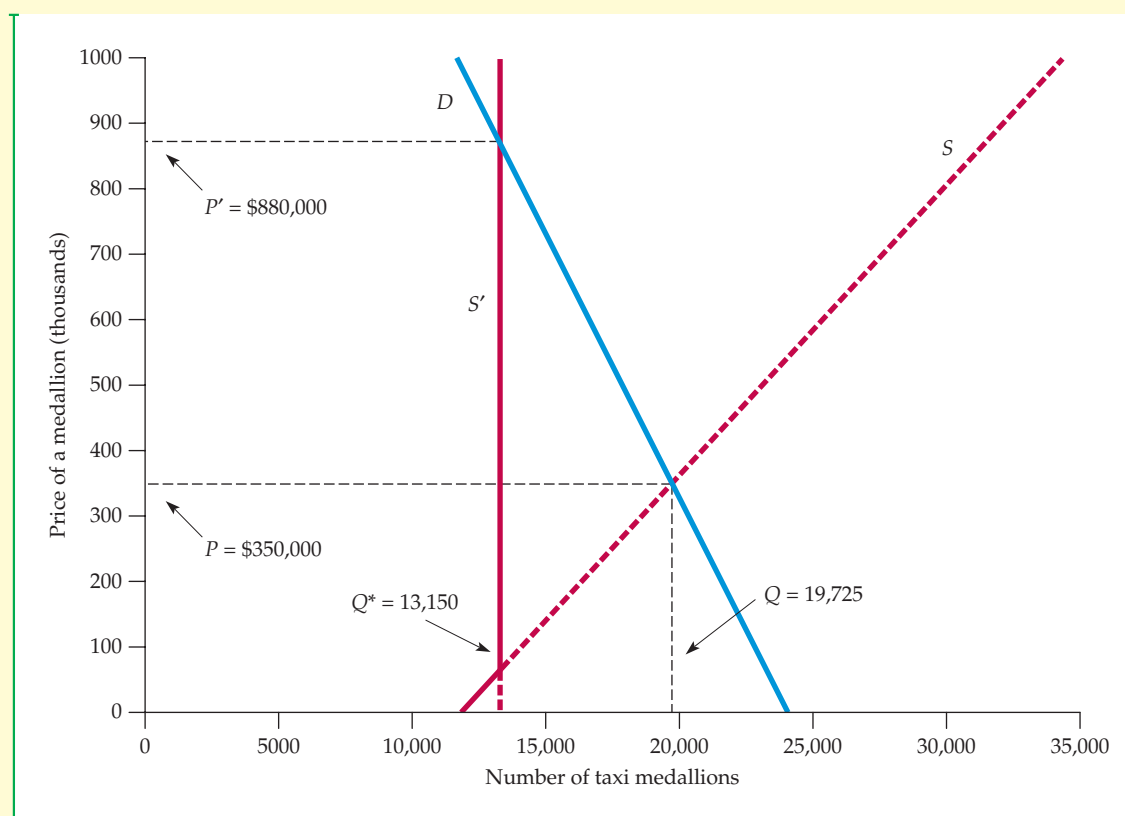


FIGURE 3.13
TAXI MEDALLIONS IN NEW YORK CITY

The demand curve D shows the quantity of medallions demanded by taxi companies as a function of the price of a medallion. The supply curve S shows the number of medallions that would be sold by current owners as a function of price. New York limits the quantity to 13,150, so the supply curve becomes vertical and intersects demand at \$880,000, the market price of a medallion in 2011.

city were to issue another 7,000 medallions for a total of about 20,000, demand and supply would equilibrate at a price of about \$350,000 per medallion – still a lot, but just enough to lease cabs, run a taxi business, and still make a profit. But supply is constrained at 13,150, at which point the supply curve (labeled S') becomes vertical, and intersects the demand curve at a price of \$880,000.

Keep in mind that New York's medallion policy hurts taxi drivers as well as citizens who depend on taxis. Most of the medallions are owned by large taxi companies—not by drivers, who must lease them from the companies (a small portion are reserved for owner-operators). To become a taxi driver, one must

take a road test and be certified. In 2011, there were 44,000 certified drivers in New York, but only 13,150 of them can drive a cab at any one time, leaving many unemployed.

Is New York City unique in its treatment of taxis? Not at all. In Boston there were only 1,825 medallions available in 2010, and medallions were bought and sold at a price of \$410,000. And just try to find a taxi in Milan, Rome, or almost any other Italian city. The Italian government severely constrains the numbers of medallions, which are owned not by large taxi companies as in New York, but by individual families, who have the political clout to preserve the value of their precious medallions.



3.5 Import Quotas and Tariffs

• **import quota** Limit on the quantity of a good that can be imported.

• **tariff** Tax on an imported good.

Many countries use **import quotas** and **tariffs** to keep the domestic price of a product above world levels and thereby enable the domestic industry to enjoy higher profits than it would under free trade. As we will see, the cost to taxpayers from this protection can be high, with the loss to consumers exceeding the gain to domestic producers.

Without a quota or tariff, a country will import a good when its world price is below the price that would prevail domestically were there no imports. Figure 9.14 illustrates this principle. S and D are the domestic supply and demand curves. If there were no imports, the domestic price and quantity would be P_0 and Q_0 , which equate supply and demand. But because the world price P_w is below P_0 , domestic consumers have an incentive to purchase from abroad and will do so if imports are not restricted. How much will be imported? The domestic price will fall to the world price P_w ; at this lower price, domestic production will fall to Q_s , and domestic consumption will rise to Q_d . Imports are then the difference between domestic consumption and domestic production, $Q_d - Q_s$.

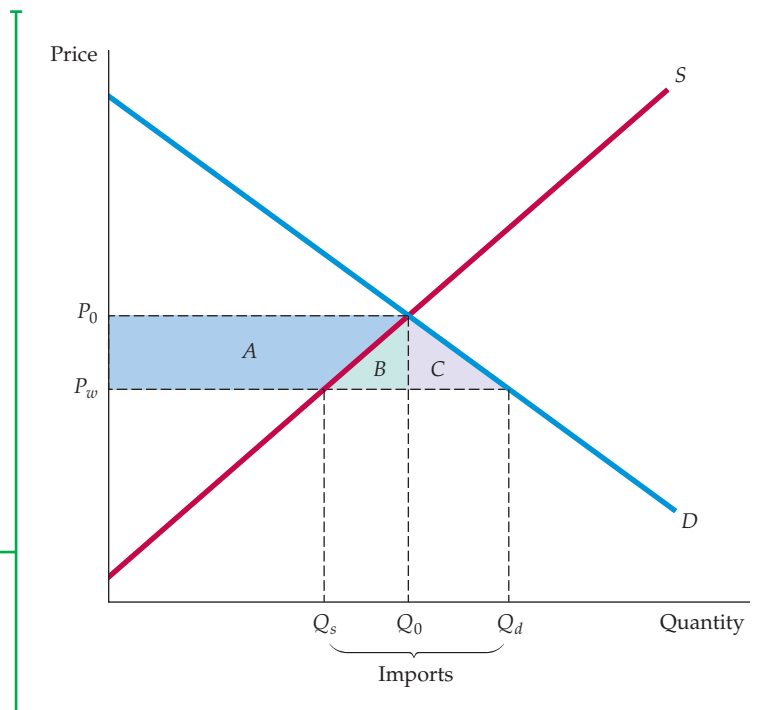
Now suppose the government, bowing to pressure from the domestic industry, eliminates imports by imposing a quota of zero—that is, forbidding any importation of the good. What are the gains and losses from such a policy?

With no imports allowed, the domestic price will rise to P_0 . Consumers who still purchase the good (in quantity Q_0) will pay more and will lose an amount of surplus given by trapezoid A and triangle B . In addition, given this higher price, some consumers will no longer buy the good, so there is an additional loss of consumer surplus, given by triangle C . The total change in consumer surplus is therefore

$$\Delta CS = -A - B - C$$

FIGURE 3.14
IMPORT TARIFF OR QUOTA THAT ELIMINATES IMPORTS

In a free market, the domestic price equals the world price P_w . A total Q_d is consumed, of which Q_s is supplied domestically and the rest imported. When imports are eliminated, the price is increased to P_0 . The gain to producers is trapezoid A . The loss to consumers is $A + B + C$, so the deadweight loss is $B + C$.





What about producers? Output is now higher (Q_0 instead of Q_s) and is sold at a higher price (P_0 instead of P_w). Producer surplus therefore increases by the amount of trapezoid A:

$$\Delta PS = A$$

The change in total surplus, $\Delta CS + \Delta PS$, is therefore $-B - C$. Again, there is a deadweight loss—consumers lose more than producers gain.

Imports could also be reduced to zero by imposing a sufficiently large tariff. The tariff would have to be equal to or greater than the difference between P_0 and P_w . With a tariff of this size, there will be no imports and, therefore, no government revenue from tariff collections, so the effect on consumers and producers would be the same as with a quota.

More often, government policy is designed to reduce but not eliminate imports. Again, this can be done with either a tariff or a quota, as Figure 9.15 shows. Under free trade, the domestic price will equal the world price P_w , and imports will be $Q_d - Q_s$. Now suppose that a tariff of T dollars per unit is imposed on imports. Then the domestic price will rise to P^* (the world price plus the tariff); domestic production will rise and domestic consumption will fall.

In Figure 9.15, this tariff leads to a change of consumer surplus given by

$$\Delta CS = -A - B - C - D$$

The change in producer surplus is again

$$\Delta PS = A$$

Finally, the government will collect revenue in the amount of the tariff times the quantity of imports, which is rectangle D . The total change in welfare, ΔCS plus ΔPS plus the revenue to the government, is therefore $-A - B - C - D + A + D = -B - C$. Triangles B and C again represent the deadweight loss from restricting

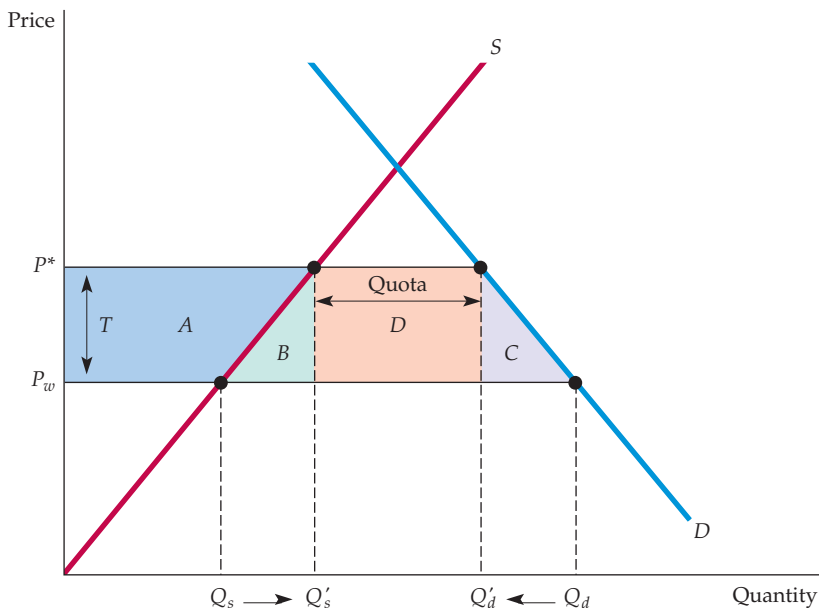


FIGURE 3.15
IMPORT TARIFF OR QUOTA
(GENERAL CASE)

When imports are reduced, the domestic price is increased from P_w to P^* . This can be achieved by a quota, or by a tariff $T = P^* - P_w$. Trapezoid A is again the gain to domestic producers. The loss to consumers is $A + B + C + D$. If a tariff is used, the government gains D , the revenue from the tariff, so the net domestic loss is $B + C$. If a quota is used instead, rectangle D becomes part of the profits of foreign producers, and the net domestic loss is $B + C + D$.



imports. (B represents the loss from domestic overproduction and C the loss from too little consumption.)

Suppose the government uses a quota instead of a tariff to restrict imports: Foreign producers can only ship a specific quantity ($Q'_d - Q'_s$ in Figure 9.15) to the United States and can then charge the higher price P^* for their U.S. sales. The changes in U.S. consumer and producer surplus will be the same as with the tariff, but instead of the U.S. government collecting the revenue given by rectangle D , this money will go to the foreign producers in the form of higher profits. The United States as a whole will be even worse off than it was under the tariff, losing D as well as the deadweight loss B and C .¹²

This situation is exactly what transpired with automobile imports from Japan in the 1980s. Under pressure from domestic automobile producers, the Reagan administration negotiated “voluntary” import restraints, under which the Japanese agreed to restrict shipments of cars to the United States. The Japanese could therefore sell those cars that were shipped at a price higher than the world level and capture a higher profit margin on each one. The United States would have been better off by simply imposing a tariff on these imports.

EXAMPLE 3.6 THE SUGAR QUOTA



In recent years, the world price of sugar has been between 10 and 28 cents per pound, while the U.S. price has been 30 to 40 cents per pound. Why? By restricting imports, the U.S. government protects the \$4 billion domestic sugar industry, which would virtually be put out of business if it had to compete with low-cost foreign producers.

This policy has been good for U.S. sugar producers. It has even been good for some foreign sugar producers—in particular, those whose successful lobbying efforts have given them big shares of the quota. But like most policies of this sort, it has been bad for consumers.

To see just how bad, let's look at the sugar market in 2010. Here are the relevant data for that year:

U.S. production:	15.9 billion pounds
U.S. consumption:	22.8 billion pounds
U.S. price:	36 cents per pound
World price:	24 cents per pound

¹²Alternatively, an import quota can be maintained by rationing imports to U.S. importing firms or trading companies. These middlemen would have the rights to import a fixed amount of the good each year. These rights are valuable because the middleman can buy the product on the world market at price P_w and then sell it at price P^* . The aggregate value of these rights is, therefore, given by rectangle D . If the government *sells* the rights for this amount of money, it can capture the same revenue it would receive with a tariff. But if these rights are given away, as sometimes happens, the money becomes a windfall to middlemen.



At these prices and quantities, the price elasticity of U.S. supply is 1.5, and the price elasticity of U.S. demand is -0.3 .¹³

We will fit linear supply and demand curves to these data, and then use them to calculate the effects of the quotas. You can verify that the following U.S. supply curve is consistent with a production level of 15.9 billion pounds, a price of 36 cents per pound, and a supply elasticity of 1.5:

$$\text{U.S. supply: } Q_S = -7.95 + 0.66P$$

where quantity is measured in billions of pounds and price in cents per pound. Similarly, the -0.3 demand elasticity, together with the data for U.S. consumption and U.S. price, give the following linear demand curve:

$$\text{U.S. demand: } Q_D = 29.73 - 0.19P$$

These supply and demand curves are plotted in Figure 9.16. Using the U.S. supply and demand curves given above, you can check that at the 24-cent world price, U.S. production would have been only about 7.9 billion pounds and U.S. consumption about 25.2 billion pounds, of which $25.2 - 7.9 = 17.3$ billion pounds would have been imported. But fortunately for U.S. producers, imports were limited to only 6.9 billion pounds.

What did limit on imports do to the U.S. price? To find out, use the U.S. supply and demand equations, and set the quantity demanded minus the quantity supplied to 6.9:

$$Q_S - Q_D = (29.73 - 0.19P) - (-7.95 + 0.66P) = 6.9$$

You can check that the solution to this equation is $P = 36.2$ cents. Thus the limit on imports pushed the U.S. price up to about 36 cents, as shown in the figure.

What did this policy cost U.S. consumers? The lost consumer surplus is given by the sum of trapezoid *A*, triangles *B* and *C*, and rectangle *D*. You should go through the calculations to verify that trapezoid *A* is equal to \$1431 million, triangle *B* to \$477 million, triangle *C* to \$137 million, and rectangle *D* to \$836 million. The total cost to consumers in 2010 was about \$2.9 billion.

How much did producers gain from this policy? Their increase in surplus is given by trapezoid *A* (i.e., about \$1.4 billion). The \$836 million of rectangle *D* was a gain for those foreign producers who succeeded in obtaining large allotments of the quota because they received a higher

In §2.6, we explain how to fit linear supply and demand functions to data of this kind.

¹³Prices and quantities are from the USDA's Economic Research Service. Find more information at <http://www.ers.usda.gov/Briefing/Sugar/Data.htm>. The elasticity estimates are based on Morris E. Morkre and David G. Tarr, *Effects of Restrictions on United States Imports: Five Case Studies and Theory*, U.S. Federal Trade Commission Staff Report, June 1981; and F. M. Scherer, "The United States Sugar Program," Kennedy School of Government Case Study, Harvard University, 1992. For a general discussion of sugar quotas and other aspects of U.S. agricultural policy, see D. Gale Johnson, *Agricultural Policy and Trade* (New York: New York University Press, 1985); and Gail L. Cramer and Clarence W. Jensen, *Agricultural Economics and Agribusiness* (New York: Wiley, 1985).

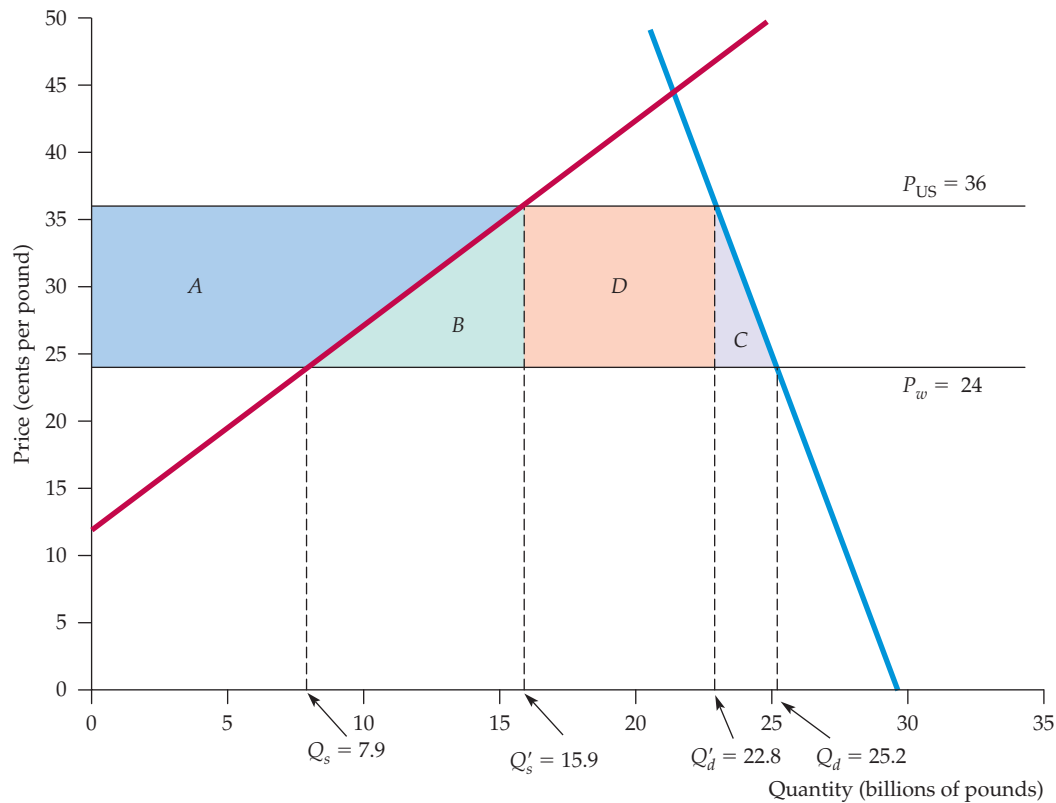


FIGURE 3.16
SUGAR QUOTA IN 2010

At the world price of 24 cents per pound, about 25.2 billion pounds of sugar would have been consumed in the United States in 2010, of which all but 7.9 billion pounds would have been imported. Restricting imports to 6.9 billion pounds caused the U.S. price to go up by 12 cents. The cost to consumers, $A + B + C + D$, was about \$2.9 billion. The gain to domestic producers was trapezoid A , about \$1.4 billion. Rectangle D , \$836 million, was a gain to those foreign producers who obtained quota allotments. Triangles B and C represent the deadweight loss of about \$614 million.

price for their sugar. Triangles B and C represent a deadweight loss of about \$614 million.

The world price of sugar has been volatile over the past decade. In the mid-2000s, the European Union removed protections on European sugar, causing the region to go from being a net sugar exporter to a net importer. Meanwhile, demand for sugar in rapidly industrializing countries like India, Pakistan and China has skyrocketed. Sugar production in these three countries is often unpredictable: while they are often net exporters, changing governmental policies and volatile weather frequently lead to decreased output, forcing them to import sugar to meet domestic demand. In addition, many countries, like Brazil, also use sugar to make ethanol, further decreasing the amount available for food.



3.6 The Impact of a Tax or Subsidy

What would happen to the price of widgets if the government imposed a \$1 tax on every widget sold? Many people would answer that the price would increase by a dollar, with consumers now paying a dollar more per widget than they would have paid without the tax. But this answer is wrong.

Or consider the following question. The government wants to impose a 50-cent-per-gallon tax on gasoline and is considering two methods of collecting it. Under Method 1, the owner of each gas station would deposit the tax money (50 cents times the number of gallons sold) in a locked box, to be collected by a government agent. Under Method 2, the buyer would pay the tax (50 cents times the number of gallons purchased) directly to the government. Which method costs the buyer more? Many people would say Method 2, but this answer is also wrong.

The burden of a tax (or the benefit of a subsidy) falls partly on the consumer and partly on the producer. Furthermore, it does not matter who puts the money in the collection box (or sends the check to the government)—Methods 1 and 2 both cost the consumer the same amount of money. As we will see, the share of a tax borne by consumers depends on the shapes of the supply and demand curves and, in particular, on the relative elasticities of supply and demand. As for our first question, a \$1 tax on widgets would indeed cause the price to rise, but usually by *less* than a dollar and sometimes by *much* less. To understand why, let's use supply and demand curves to see how consumers and producers are affected when a tax is imposed on a product, and what happens to price and quantity.

THE EFFECTS OF A SPECIFIC TAX For the sake of simplicity, we will consider a **specific tax**—a tax of a certain amount of money *per unit sold*. This is in contrast to an *ad valorem* (i.e., proportional) tax, such as a state sales tax. (The analysis of an *ad valorem* tax is roughly the same and yields the same qualitative results.) Examples of specific taxes include federal and state taxes on gasoline and cigarettes.

• **specific tax** Tax of a certain amount of money per unit sold.

Suppose the government imposes a tax of t cents per unit on widgets. Assuming that everyone obeys the law, the government must then receive t cents for every widget sold. *This means that the price the buyer pays must exceed the net price the seller receives by t cents.* Figure 9.17 illustrates this simple accounting relationship—and its implications. Here, P_0 and Q_0 represent the market price and quantity *before* the tax is imposed. P_b is the price that buyers pay, and P_s is the net price that sellers receive *after* the tax is imposed. Note that $P_b - P_s = t$, so the government is happy.

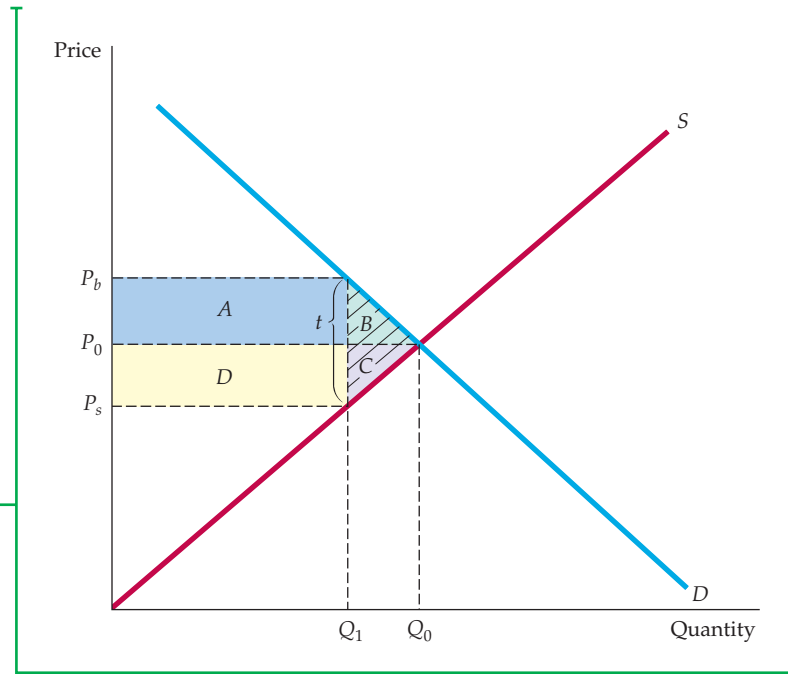
How do we determine what the market quantity will be after the tax is imposed, and how much of the tax is borne by buyers and how much by sellers? First, remember that what buyers care about is the price that they must pay: P_b . The amount that they will buy is given by the demand curve; it is the quantity that we read off of the demand curve given a price P_b . Similarly, sellers care about the net price they receive, P_s . Given P_s , the quantity that they will produce and sell is read off the supply curve. Finally, we know that the quantity sold must equal the quantity bought. The solution, then, is to find the quantity that corresponds to a price of P_b on the demand curve, and a price of P_s on the supply curve, such that the difference $P_b - P_s$ is equal to the tax t . In Figure 9.17, this quantity is shown as Q_1 .

Who bears the burden of the tax? In Figure 9.17, this burden is shared roughly equally by buyers and sellers. The market price (the price buyers pay) rises by half of the tax, and the price that sellers receive falls by roughly half of the tax.



FIGURE 3.17 INCIDENCE OF A TAX

P_b is the price (including the tax) paid by buyers. P_s is the price that sellers receive, less the tax. Here the burden of the tax is split evenly between buyers and sellers. Buyers lose $A + B$, sellers lose $D + C$, and the government earns $A + D$ in revenue. The deadweight loss is $B + C$.



As Figure 9.17 shows, market clearing requires *four conditions* to be satisfied after the tax is in place:

1. The quantity sold and the buyer's price P_b must lie on the demand curve (because buyers are interested only in the price they must pay).
2. The quantity sold and the seller's price P_s must lie on the supply curve (because sellers are concerned only with the amount of money they receive net of the tax).
3. The quantity demanded must equal the quantity supplied (Q_1 in the figure).
4. The difference between the price the buyer pays and the price the seller receives must equal the tax t .

These conditions can be summarized by the following four equations:

$$Q^D = Q^D(P_b) \quad (9.1a)$$

$$Q^S = Q^S(P_s) \quad (9.1b)$$

$$Q^D = Q^S \quad (9.1c)$$

$$P_b - P_s = t \quad (9.1d)$$

If we know the demand curve $Q^D(P_b)$, the supply curve $Q^S(P_s)$, and the size of the tax t , we can solve these equations for the buyers' price P_b , the sellers' price P_s , and the total quantity demanded and supplied. This task is not as difficult as it may seem, as we will demonstrate in Example 9.7.

Figure 9.17 also shows that a tax results in a *deadweight loss*. Because buyers pay a higher price, there is a change in consumer surplus given by

$$\Delta CS = -A - B$$



Because sellers now receive a lower price, there is also a change in producer surplus given by

$$\Delta PS = -C - D$$

Government tax revenue is tQ_1 , the sum of rectangles A and D . The total change in welfare, ΔCS plus ΔPS plus the revenue to the government, is therefore $-A - B - C - D + A + D = -B - C$. Triangles B and C represent the deadweight loss from the tax.

In Figure 9.17, the burden of the tax is shared almost evenly between buyers and sellers, but this is not always the case. If demand is relatively inelastic and supply is relatively elastic, the burden of the tax will fall mostly on buyers. Figure 9.18(a) shows why: It takes a relatively large increase in price to reduce the quantity demanded by even a small amount, whereas only a small price decrease is needed to reduce the quantity supplied. For example, because cigarettes are addictive, the elasticity of demand is small (about -0.4); thus federal and state cigarette taxes are borne largely by cigarette buyers.¹⁴

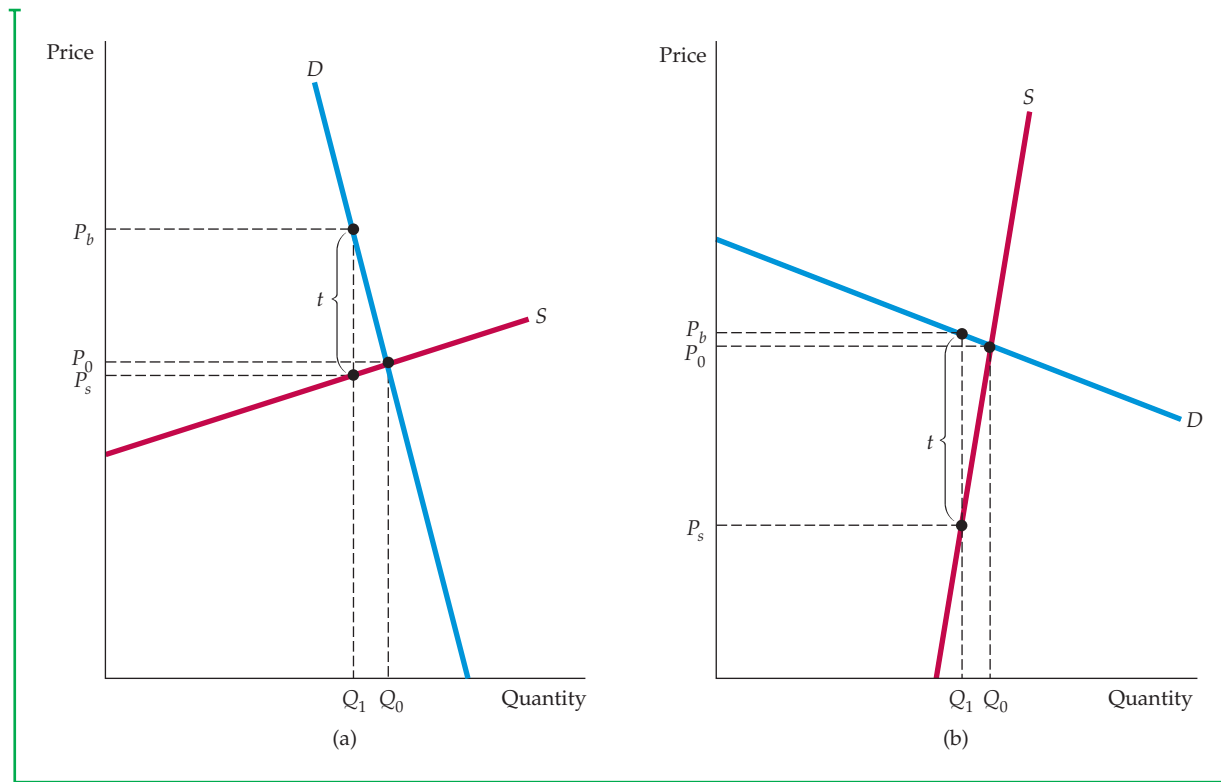


FIGURE 3.18
IMPACT OF A TAX DEPENDS ON ELASTICITIES OF SUPPLY AND DEMAND

- (a) If demand is very inelastic relative to supply, the burden of the tax falls mostly on buyers.
(b) If demand is very elastic relative to supply, it falls mostly on sellers.

¹⁴See Daniel A. Sumner and Michael K. Wohlgenant, "Effects of an Increase in the Federal Excise Tax on Cigarettes," *American Journal of Agricultural Economics* 67 (May 1985): 235–42.



Figure 9.18(b) shows the opposite case: If demand is relatively elastic and supply is relatively inelastic, the burden of the tax will fall mostly on sellers.

So even if we have only estimates of the elasticities of demand and supply at a point or for a small range of prices and quantities, instead of the entire demand and supply curves, we can still roughly determine who will bear the greatest burden of a tax (whether the tax is actually in effect or is only under discussion as a policy option). In general, *a tax falls mostly on the buyer if E_d/E_s is small, and mostly on the seller if E_d/E_s is large.*

In fact, by using the following “pass-through” formula, we can calculate the percentage of the tax borne by buyers:

$$\text{Pass-through fraction} = E_s / (E_s - E_d)$$

This formula tells us what fraction of the tax is “passed through” to consumers in the form of higher prices. For example, when demand is totally inelastic, so that E_d is zero, the pass-through fraction is 1 and all the tax is borne by consumers. When demand is totally elastic, the pass-through fraction is zero and producers bear all the tax. (The fraction of the tax that producers bear is given by $-E_d / (E_s - E_d)$.)

The Effects of a Subsidy

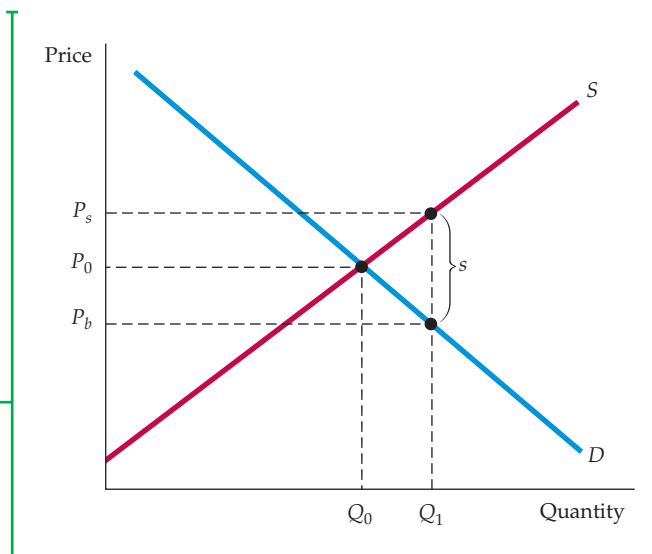
• **subsidy** Payment reducing the buyer's price below the seller's price; i.e., a negative tax.

A **subsidy** can be analyzed in much the same way as a tax—in fact, you can think of a subsidy as a *negative tax*. With a subsidy, the sellers' price *exceeds* the buyers' price, and the difference between the two is the amount of the subsidy. As you would expect, the effect of a subsidy on the quantity produced and consumed is just the opposite of the effect of a tax—the quantity will increase.

Figure 9.19 illustrates this. At the presubsidy market price P_0 , the elasticities of supply and demand are roughly equal. As a result, the benefit of the subsidy is shared roughly equally between buyers and sellers. As with a tax, this is not always the case. In general, *the benefit of a subsidy accrues mostly to buyers if E_d/E_s is small and mostly to sellers if E_d/E_s is large.*

FIGURE 3.19 SUBSIDY

A subsidy can be thought of as a negative tax. Like a tax, the benefit of a subsidy is split between buyers and sellers, depending on the relative elasticities of supply and demand.





As with a tax, given the supply curve, the demand curve, and the size of the subsidy s , we can solve for the resulting prices and quantity. The same four conditions needed for the market to clear apply for a subsidy as for a tax, but now the difference between the sellers' price and the buyers' price is equal to the subsidy. Again, we can write these conditions algebraically:

$$Q^D = Q^D(P_b) \quad (9.2a)$$

$$Q^S = Q^S(P_s) \quad (9.2b)$$

$$Q^D = Q^S \quad (9.2c)$$

$$P_s - P_b = s \quad (9.2d)$$

To make sure you understand how to analyze the impact of a tax or subsidy, you might find it helpful to work through one or two examples, such as Exercises 2 and 14 at the end of this chapter.

In §2.5, we explain that demand is often more price elastic in the long run than in the short run because it takes time for people to change their consumption habits and/or because the demand for a good might be linked to the stock of another good that changes slowly.

EXAMPLE 3.7 A TAX ON GASOLINE

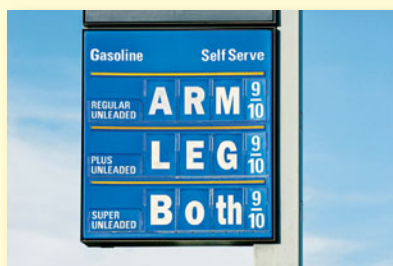
The idea of a large tax on gasoline, both to raise government revenue and to reduce oil consumption and U.S. dependence on oil imports, has been discussed for many years. Let's see how a \$1.00-per-gallon tax would affect the price and consumption of gasoline.

We will do this analysis in the setting of market conditions during 2005–2010—when gasoline was selling for about \$2 per gallon on average and total consumption was about 100 billion gallons per year (bg/yr).¹⁵ We will also use intermediate-run elasticities: elasticities that would apply to a period of about three to six years after a price change.

A reasonable number for the intermediate-run elasticity of gasoline demand is -0.5 (see Example 2.6 in Chapter 2—page 43). We can use this figure, together with the \$2 and 100 bg/yr price and quantity numbers, to calculate a linear demand curve for gasoline. You can verify that the following demand curve fits these data:

$$\text{Gasoline demand: } Q^D = 150 - 25P$$

Gasoline is refined from crude oil, some of which is produced domestically and some imported. (Some gasoline is also imported directly.) The supply curve for gasoline will therefore depend on the world price of oil, on domestic oil supply, and on the cost of refining. The details are beyond the scope of this example, but a reasonable number for the elasticity of supply is 0.4 . You should verify that this elasticity, together with



For a review of the procedure for calculating linear curves, see §2.6. Given data for price and quantity, as well as estimates of demand and supply elasticities, we can use a two-step procedure to solve for quantity demanded and supplied.

¹⁵Of course, this price varied across regions and grades of gasoline, but we can ignore this here. Quantities of oil and oil products are often measured in barrels; there are 42 gallons in a barrel, so the quantity figure could also be written as 2.4 billion barrels per year.



the \$2 and 100 bg/yr price and quantity, gives the following linear supply curve:

$$\text{Gasoline supply: } Q^S = 60 + 20P$$

You should also verify that these demand and supply curves imply a market price of \$2 and quantity of 100 bg/yr.

We can use these linear demand and supply curves to calculate the effect of a \$1-per-gallon tax. First, we write the four conditions that must hold, as given by equations (9.2a–d):

$$Q^D = 150 - 25P_b \quad (\text{Demand})$$

$$Q^S = 60 + 20P_s \quad (\text{Supply})$$

$$Q^D = Q^S \quad (\text{Supply must equal demand})$$

$$P_b = P_s = 1.00 \quad (\text{Government must receive \$1.00/gallon})$$

Now combine the first three equations to equate supply and demand:

$$150 - 25P_b = 60 + 20P_s$$

We can rewrite the last of the four equations as $P_b = P_s + 1.00$ and substitute this for P_b in the above equation:

$$150 - 25(P_s + 1.00) = 60 + 20P_s$$

Now we can rearrange this equation and solve for P_s :

$$20P_s + 25P_s = 150 - 25 - 60$$

$$45P_s = 65, \text{ or } P_s = 1.44$$

Remember that $P_b = P_s + 1.00$, so $P_b = 2.44$. Finally, we can determine the total quantity from either the demand or supply curve. Using the demand curve (and the price $P_b = 2.44$), we find that $Q = 150 - (25)(2.44) = 150 - 61$, or $Q = 89$ bg/yr. This represents an 11-percent decline in gasoline consumption. Figure 9.20 illustrates these calculations and the effect of the tax.

The burden of this tax would be split roughly evenly between consumers and producers. Consumers would pay about 44 cents per gallon more for gasoline, and producers would receive about 56 cents per gallon less. It should not be surprising, then, that both consumers and producers opposed such a tax, and politicians representing both groups fought the proposal every time it came up. But note that the tax would raise significant revenue for the government. The annual revenue would be $tQ = (1.00)(89) = \$89$ billion per year.

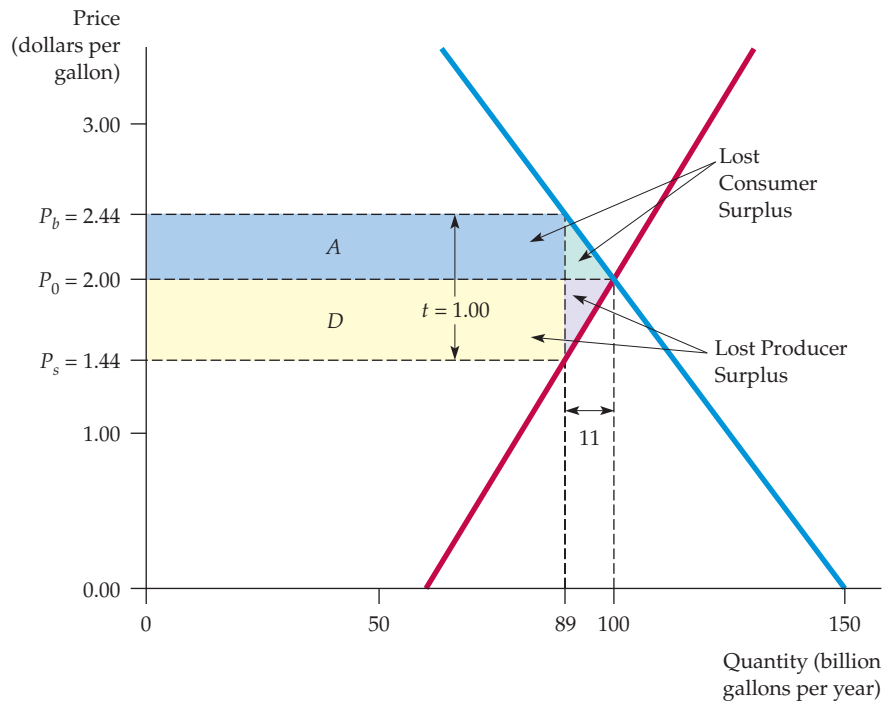
The cost to consumers and producers, however, will be more than the \$89 billion in tax revenue. Figure 9.20 shows the deadweight loss from this tax as the two shaded triangles. The two rectangles A and D represent the total tax collected by the government, but the total loss of consumer and producer surplus is larger.

Before deciding whether a gasoline tax is desirable, it is important to know how large the resulting deadweight loss is likely to be. We can easily



FIGURE 3.20
IMPACT OF \$1
GASOLINE TAX

The price of gasoline at the pump increases from \$2.00 per gallon to \$2.44, and the quantity sold falls from 100 to 89 bg/yr. Annual revenue from the tax is $(1.00)(89) = \$89$ billion. The two triangles show the deadweight loss of \$5.5 billion per year.



calculate this from Figure 9.20. Combining the two small triangles into one large one, we see that the area is

$$\begin{aligned} & (1/2) \times (\$1.00/\text{gallon}) \times (11 \text{ billion gallons/year}) \\ & = \$5.5 \text{ billion per year} \end{aligned}$$

This deadweight loss is about 6 percent of the government revenue resulting from the tax, and must be balanced against any additional benefits that the tax might bring.

SUMMARY

1. Simple models of supply and demand can be used to analyze a wide variety of government policies, including price controls, minimum prices, price support programs, production quotas or incentive programs to limit output, import tariffs and quotas, and taxes and subsidies.
2. In each case, consumer and producer surplus are used to evaluate the gains and losses to consumers and producers. Applying the methodology to natural gas price controls, airline regulation, price supports for wheat, and the sugar quota shows that these gains and losses can be quite large.
3. When government imposes a tax or subsidy, price usually does not rise or fall by the full amount of the tax or subsidy. Also, the incidence of a tax or subsidy is usually split between producers and consumers. The fraction that each group ends up paying or receiving depends on the relative elasticities of supply and demand.
4. Government intervention generally leads to a deadweight loss; even if consumer surplus and producer surplus are weighted equally, there will be a net loss from government policies that shifts surplus from one group to the other. In some cases, this deadweight loss



will be small, but in other cases—price supports and import quotas are examples—it is large. This deadweight loss is a form of economic inefficiency that must be taken into account when policies are designed and implemented.

5. Government intervention in a competitive market is not always bad. Government—and the society it

represents—might have objectives other than economic efficiency. There are also situations in which government intervention can improve economic efficiency. Examples are externalities and cases of market failure. These situations, and the way government can respond to them, are discussed in Chapters 17 and 18.

QUESTIONS FOR REVIEW

1. What is meant by *deadweight loss*? Why does a price ceiling usually result in a deadweight loss?
2. Suppose the supply curve for a good is completely inelastic. If the government imposed a price ceiling below the market-clearing level, would a deadweight loss result? Explain.
3. How can a price ceiling make consumers better off? Under what conditions might it make them worse off?
4. Suppose the government regulates the price of a good to be no lower than some minimum level. Can such a minimum price make producers as a whole worse off? Explain.
5. How are production limits used in practice to raise the prices of the following goods or services: (a) taxi rides, (b) drinks in a restaurant or bar, (c) wheat or corn?
6. Suppose the government wants to increase farmers' incomes. Why do price supports or acreage-limitation programs cost society more than simply giving farmers money?
7. Suppose the government wants to limit imports of a certain good. Is it preferable to use an import quota or a tariff? Why?
8. The burden of a tax is shared by producers and consumers. Under what conditions will consumers pay most of the tax? Under what conditions will producers pay most of it? What determines the share of a subsidy that benefits consumers?
9. Why does a tax create a deadweight loss? What determines the size of this loss?

EXERCISES

1. From time to time, Congress has raised the minimum wage. Some people suggested that a government subsidy could help employers finance the higher wage. This exercise examines the economics of a minimum wage and wage subsidies. Suppose the supply of low-skilled labor is given by
2. Suppose the market for widgets can be described by the following equations:

$$\text{Demand: } P = 10 - Q$$

$$\text{Supply: } P = Q - 4$$

where L^S is the quantity of low-skilled labor (in millions of persons employed each year), and w is the wage rate (in dollars per hour). The demand for labor is given by

$$L^D = 80 - 10w$$

- a. What will be the free-market wage rate and employment level? Suppose the government sets a minimum wage of \$5 per hour. How many people would then be employed?
- b. Suppose that instead of a minimum wage, the government pays a subsidy of \$1 per hour for each employee. What will the total level of employment be now? What will the equilibrium wage rate be?

where P is the price in dollars per unit and Q is the quantity in thousands of units. Then:

- a. What is the equilibrium price and quantity?
 - b. Suppose the government imposes a tax of \$1 per unit to reduce widget consumption and raise government revenues. What will the new equilibrium quantity be? What price will the buyer pay? What amount per unit will the seller receive?
 - c. Suppose the government has a change of heart about the importance of widgets to the happiness of the American public. The tax is removed and a subsidy of \$1 per unit granted to widget producers. What will the equilibrium quantity be? What price will the buyer pay? What amount per unit (including the subsidy) will the seller receive? What will be the total cost to the government?
3. Japanese rice producers have extremely high production costs, due in part to the high opportunity cost of



land and to their inability to take advantage of economies of large-scale production. Analyze two policies intended to maintain Japanese rice production: (1) a per-pound subsidy to farmers for each pound of rice produced, or (2) a per-pound tariff on imported rice. Illustrate with supply-and-demand diagrams the equilibrium price and quantity, domestic rice production, government revenue or deficit, and deadweight loss from each policy. Which policy is the Japanese government likely to prefer? Which policy are Japanese farmers likely to prefer?

4. In 1983, the Reagan administration introduced a new agricultural program called the Payment-in-Kind Program. To see how the program worked, let's consider the wheat market:

- Suppose the demand function is $Q^D = 28 - 2P$ and the supply function is $Q^S = 4 + 4P$, where P is the price of wheat in dollars per bushel, and Q is the quantity in billions of bushels. Find the free-market equilibrium price and quantity.
- Now suppose the government wants to lower the supply of wheat by 25 percent from the free-market equilibrium by paying farmers to withdraw land from production. However, the payment is made in wheat rather than in dollars—hence the name of the program. The wheat comes from vast government reserves accumulated from previous price support programs. The amount of wheat paid is equal to the amount that could have been harvested on the land withdrawn from production. Farmers are free to sell this wheat on the market. How much is now produced by farmers? How much is indirectly supplied to the market by the government? What is the new market price? How much do farmers gain? Do consumers gain or lose?
- Had the government not given the wheat back to the farmers, it would have stored or destroyed it. Do taxpayers gain from the program? What potential problems does the program create?

5. About 100 million pounds of jelly beans are consumed in the United States each year, and the price has been about 50 cents per pound. However, jelly bean producers feel that their incomes are too low and have convinced the government that price supports are in order. The government will therefore buy up as many jelly beans as necessary to keep the price at \$1 per pound. However, government economists are worried about the impact of this program because they have no estimates of the elasticities of jelly bean demand or supply.

- Could this program cost the government *more* than \$50 million per year? Under what conditions? Could it cost *less* than \$50 million per year? Under what conditions? Illustrate with a diagram.
- Could this program cost consumers (in terms of lost consumer surplus) *more* than \$50 million per year? Under what conditions? Could it

cost consumers *less* than \$50 million per year? Under what conditions? Again, use a diagram to illustrate.

6. In Exercise 4 in Chapter 2 (page 62), we examined a vegetable fiber traded in a competitive world market and imported into the United States at a world price of \$9 per pound. U.S. domestic supply and demand for various price levels are shown in the following table.

PRICE	U.S. SUPPLY (MILLION POUNDS)	U.S. DEMAND (MILLION POUNDS)
3	2	34
6	4	28
9	6	22
12	8	16
15	10	10
18	12	4

Answer the following questions about the U.S. market:

- Confirm that the demand curve is given by $Q_D = 40 - 2P$, and that the supply curve is given by $Q_S = 2/3P$.
 - Confirm that if there were no restrictions on trade, the United States would import 16 million pounds.
 - If the United States imposes a tariff of \$3 per pound, what will be the U.S. price and level of imports? How much revenue will the government earn from the tariff? How large is the deadweight loss?
 - If the United States has no tariff but imposes an import quota of 8 million pounds, what will be the U.S. domestic price? What is the cost of this quota for U.S. consumers of the fiber? What is the gain for U.S. producers?
7. The United States currently imports all of its coffee. The annual demand for coffee by U.S. consumers is given by the demand curve $Q = 250 - 10P$, where Q is quantity (in millions of pounds) and P is the market price per pound of coffee. World producers can harvest and ship coffee to U.S. distributors at a constant marginal (= average) cost of \$8 per pound. U.S. distributors can in turn distribute coffee for a constant \$2 per pound. The U.S. coffee market is competitive. Congress is considering a tariff on coffee imports of \$2 per pound.
- If there is no tariff, how much do consumers pay for a pound of coffee? What is the quantity demanded?
 - If the tariff is imposed, how much will consumers pay for a pound of coffee? What is the quantity demanded?



- c. Calculate the lost consumer surplus.
 - d. Calculate the tax revenue collected by the government.
 - e. Does the tariff result in a net gain or a net loss to society as a whole?
8. A particular metal is traded in a highly competitive world market at a world price of \$9 per ounce. Unlimited quantities are available for import into the United States at this price. The supply of this metal from domestic U.S. mines and mills can be represented by the equation $Q^S = 2/3P$, where Q^S is U.S. output in million ounces and P is the domestic price. The demand for the metal in the United States is $Q^D = 40 - 2P$, where Q^D is the domestic demand in million ounces.
- In recent years the U.S. industry has been protected by a tariff of \$9 per ounce. Under pressure from other foreign governments, the United States plans to reduce this tariff to zero. Threatened by this change, the U.S. industry is seeking a voluntary restraint agreement that would limit imports into the United States to 8 million ounces per year.
- a. Under the \$9 tariff, what was the U.S. domestic price of the metal?
 - b. If the United States eliminates the tariff and the voluntary restraint agreement is approved, what will be the U.S. domestic price of the metal?
9. Among the tax proposals regularly considered by Congress is an additional tax on distilled liquors. The tax would not apply to beer. The price elasticity of supply of liquor is 4.0, and the price elasticity of demand is -0.2 . The cross-elasticity of demand for beer with respect to the price of liquor is 0.1.
- a. If the new tax is imposed, who will bear the greater burden—liquor suppliers or liquor consumers? Why?
 - b. Assuming that beer supply is infinitely elastic, how will the new tax affect the beer market?
10. In Example 9.1 (page 322), we calculated the gains and losses from price controls on natural gas and found that there was a deadweight loss of \$5.68 billion. This calculation was based on a price of oil of \$50 per barrel.
- a. If the price of oil were \$60 per barrel, what would be the free-market price of gas? How large a deadweight loss would result if the maximum allowable price of natural gas were \$3.00 per thousand cubic feet?
 - b. What price of oil would yield a free-market price of natural gas of \$3?
11. Example 9.6 (page 342) describes the effects of the sugar quota. In 2011, imports were limited to 6.9 billion pounds, which pushed the domestic price to 36

cents per pound. Suppose imports were expanded to 10 billion pounds.

- a. What would be the new U.S. domestic price?
 - b. How much would consumers gain and domestic producers lose?
 - c. What would be the effect on deadweight loss and foreign producers?
12. The domestic supply and demand curves for hula beans are as follows:

$$\text{Supply: } P = 50 + Q$$

$$\text{Demand: } P = 200 - 2Q$$

where P is the price in cents per pound and Q is the quantity in millions of pounds. The U.S. is a small producer in the world hula bean market, where the current price (which will not be affected by anything we do) is 60 cents per pound. Congress is considering a tariff of 40 cents per pound. Find the domestic price of hula beans that will result if the tariff is imposed. Also compute the dollar gain or loss to domestic consumers, domestic producers, and government revenue from the tariff.

13. Currently, the social security payroll tax in the United States is evenly divided between employers and employees. Employers must pay the government a tax of 6.2 percent of the wages they pay, and employees must pay 6.2 percent of the wages they receive. Suppose the tax were changed so that employers paid the full 12.4 percent and employees paid nothing. Would employees be better off?
14. You know that if a tax is imposed on a particular product, the burden of the tax is shared by producers and consumers. You also know that the demand for automobiles is characterized by a stock adjustment process. Suppose a special 20-percent sales tax is suddenly imposed on automobiles. Will the share of the tax paid by consumers rise, fall, or stay the same over time? Explain briefly. Repeat for a 50-cents-per-gallon gasoline tax.
15. In 2011, Americans smoked 16 billion packs of cigarettes. They paid an average retail price of \$5.00 per pack.
- a. Given that the elasticity of supply is 0.5 and the elasticity of demand is -0.4 , derive linear demand and supply curves for cigarettes.
 - b. Cigarettes are subject to a federal tax, which was about \$1.00 per pack in 2011. What does this tax do to the market-clearing price and quantity?
 - c. How much of the federal tax will consumers pay? What part will producers pay?

CHAPTER 4

Overview of the Labor Market

Chapters 4, 5, and 6 are taken from *Modern Labor Economics: Theory and Public Policy*, Twelfth Edition by Ehrenberg and Smith

Every society—regardless of its wealth, its form of government, or the organization of its economy—must make basic decisions. It must decide what and how much to produce, how to produce it, and how the output shall be distributed. These decisions require finding out what consumers want, what technologies for production are available, and what the skills and preferences of workers are; deciding where to produce; and coordinating all such decisions so that, for example, the millions of people in New York City and the isolated few in an Alaskan fishing village can each buy the milk, bread, meat, vanilla extract, mosquito repellent, and brown shoe polish they desire at the grocery store. The process of coordination involves creating incentives so that the right amount of labor and capital will be employed at the right place at the required time.

These decisions can, of course, be made by administrators employed by a centralized bureaucracy. The amount of information this bureaucracy must obtain and process to make the millions of needed decisions wisely, and the number of incentives it must create to ensure that these decisions are coordinated, are truly mind-boggling. It boggles the mind even more to consider the major alternative to centralized decision making—the decentralized marketplace. Millions of producers striving to make a profit observe the prices millions of consumers are willing to pay for products and the wages millions of workers are willing to accept for work. Combining these pieces

of information with data on various technologies, they decide where to produce, what to produce, whom to hire, and how much to produce. No one is in charge, and while market imperfections impede progress toward achieving the best allocation of resources, millions of people find jobs that enable them to purchase the items they desire each year. The production, employment, and consumption decisions are all made and coordinated by price signals arising through the marketplace.

The market that allocates workers to jobs and coordinates employment decisions is the *labor market*. With roughly 150 million workers and almost 8 million employers in the United States, thousands of decisions about career choice, hiring, quitting, compensation, and technology must be made and coordinated every day.

Because we believe that it is essential for students to understand the “big picture” at the outset, this chapter presents an overview of what the labor market does and how it works. After seeing how the buying and selling sides of the labor market are coordinated at an overall (or “market”) level, we then turn to more detailed analyses of individual behavior on each side in subsequent chapters.

The Labor Market: Definitions, Facts, and Trends

Every market has buyers and sellers, and the labor market is no exception: the buyers are employers, and the sellers are workers. Some of these participants may not be active at any given moment in the sense of seeking new employees or new jobs, but on any given day, thousands of firms and workers will be “in the market” trying to transact. If, as in the case of doctors or mechanical engineers, buyers and sellers are searching throughout the entire nation for each other, we would describe the market as a *national labor market*. If buyers and sellers search only locally, as in the case of data entry clerks or automobile mechanics, the labor market is a *local* one.

When we speak of a particular “labor market”—for taxi drivers, say—we are using the term *loosely* to refer to the companies trying to hire people to drive their cabs and the people seeking employment as cabdrivers. The efforts of these buyers and sellers of labor to transact and establish an employment relationship constitute the market for cabdrivers. However, neither the employers nor the drivers are confined to this market; both could simultaneously be in other markets as well. An entrepreneur with \$100,000 to invest might be thinking of operating either a taxi company or a car wash, depending on the projected revenues and costs of each. A person seeking a cab-driving job might also be trying to find work as an actor. Thus, all the various labor markets that we can define on the basis of industry, occupation, geography, transaction rules, or job character are interrelated to some degree. We speak of these narrowly defined labor markets for the sake of convenience.

Some labor markets, particularly those in which the sellers of labor are represented by a union, operate under a very formal set of rules that partly govern buyer–seller transactions. In the unionized construction trades, for example,

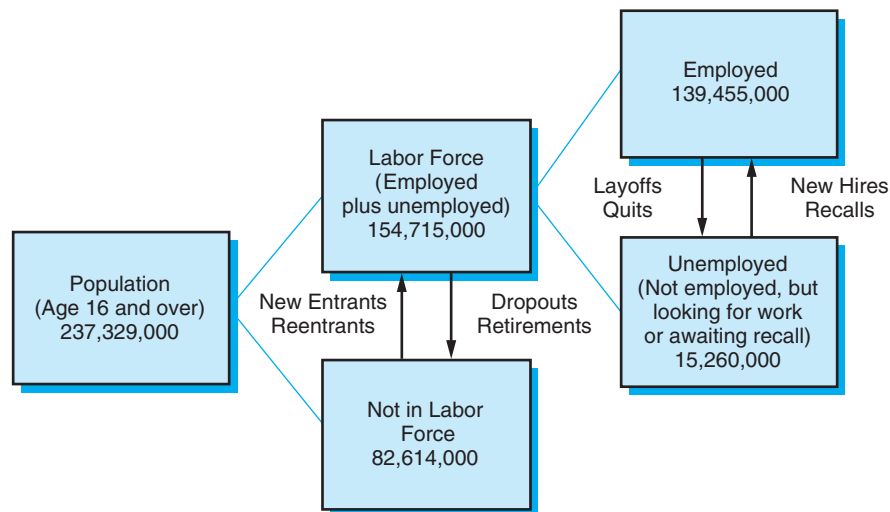
employers must hire at the union hiring hall from a list of eligible union members. In other unionized markets, the employer has discretion over who gets hired but is constrained by a union–management agreement in such matters as the order in which employees may be laid off, procedures regarding employee complaints, and promotions. The markets for government jobs and jobs with large nonunion employers also tend to operate under rules that constrain the authority of management and ensure fair treatment of employees. When a formal set of rules and procedures guides and constrains the employment relationship *within* a firm, an *internal labor market* is said to exist.¹

The Labor Force and Unemployment

Figure 2.1 highlights some basic definitions concerning labor market status. The term *labor force* refers to all those over 16 years of age who are employed, actively seeking work, or expecting recall from a layoff. Those in the labor force who are not employed for pay are the *unemployed*.² People who are not employed and are

Figure 4.1

Labor Force Status of
the U.S. Adult Civilian
Population, April 2010
(seasonally adjusted)



¹An analysis of internal labor markets can be found in Michael L. Wachter and Randall Wright, "The Economics of Internal Labor Markets," *University of Pennsylvania Law Review* 29 (Spring 1990): 240–262.

²The official definition of unemployment for purposes of government statistics includes those who have been laid off by their employers, those who have been fired or have quit and are looking for other work, and those who are just entering or reentering the labor force but have not found a job as yet. The extent of unemployment is estimated from a monthly survey of some 50,000 households called the Current Population Survey (CPS). Interviewers ascertain whether household members are employed, whether they meet one of the aforementioned conditions (in which case they are considered "unemployed"), or whether they are out of the labor force.

neither looking for work nor waiting to be recalled from layoff by their employers are not counted as part of the labor force. The total labor force thus consists of the employed and the unemployed.

The number and identities of people in each labor market category are always changing, and as we shall see in chapter 14, the flows of people from one category to another are sizable. As Figure 2.1 suggests, there are four major flows between labor market states:

1. Employed workers become unemployed by *quitting* voluntarily or *being laid off* (being involuntarily separated from the firm, either temporarily or permanently).
2. Unemployed workers obtain employment by *being newly hired* or *being recalled* to a job from which they were temporarily laid off.
3. Those in the labor force, whether employed or unemployed, can leave the labor force by *retiring* or otherwise deciding against taking or seeking work for pay (*dropping out*).
4. Those who have never worked or looked for a job expand the labor force by *entering* it, while those who have dropped out do so by *reentering* the labor force.

In April 2010, there were almost 155 million people in the labor force, representing about 66 percent of the entire population over 16 years of age. An overall *labor force participation rate* (labor force divided by population) of 65 percent is higher than the rates of about 60 percent that prevailed prior to the 1980s but—as is shown in Table 2.1—a bit lower than the rate in 2000. Underlying changes over time in the overall labor force participation rate are a continued decline in the participation rate for men and a dramatic rise in the participation rate for women

Table 4.1

Labor Force Participation Rates by Gender, 1950–2010

Year	Total (%)	Men (%)	Women (%)
1950	59.9	86.8	33.9
1960	60.2	84.0	37.8
1970	61.3	80.6	43.4
1980	64.2	77.9	51.6
1990	66.5	76.4	57.5
2000	67.2	74.7	60.2
2010 (April)	65.2	71.8	59.0

Sources: 1950–1980: U.S. President, *Employment and Training Report of the President* (Washington, D.C.: U.S. Government Printing Office), transmitted to the Congress 1981, Table A-1.

1990: U.S. Bureau of Labor Statistics, *Employment and Earnings* 45 (February 1998), Tables A-1 and A-2.

2000: U.S. Bureau of Labor Statistics, *Employment Situation* (News Release, October 2001), Table A-1.

2010: U.S. Bureau of Labor Statistics, *Employment Situation* (Economic News Release, May 2010), Table A-1.

Data and news releases are available online at <http://www.bls.gov>.

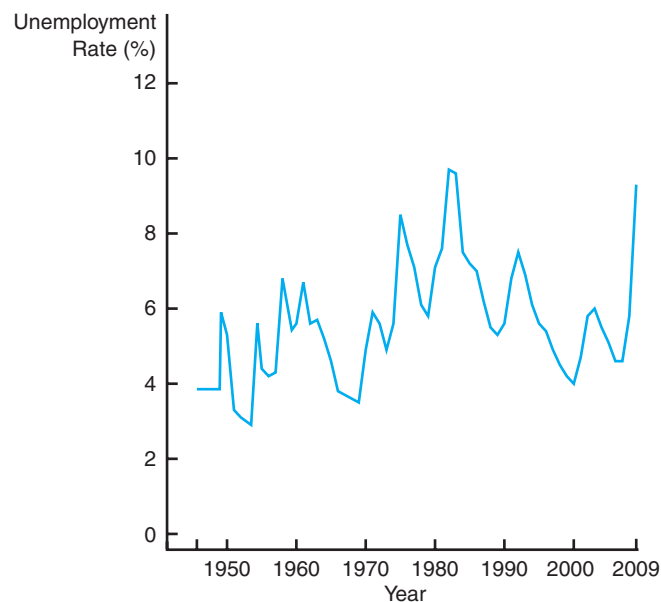
prior to 2000, with a modest decline since then. These trends and their causes will be discussed in detail in chapters 6 and 7.

The ratio of those unemployed to those in the labor force is the *unemployment rate*. While this rate is crude and has several imperfections, it is the most widely cited measure of labor market conditions. When the unemployment rate is around 5 percent in the United States, the labor market is considered *tight*, indicating that jobs in general are plentiful and hard for employers to fill and that most of those who are unemployed will find other work quickly. When the unemployment rate is higher—say, 7 percent or above—the labor market is described as *loose*, in the sense that workers are abundant and jobs are relatively easy for employers to fill. To say that the labor market as a whole is loose, however, does not imply that no shortages can be found anywhere; to say it is tight can still mean that in some occupations or places the number of those seeking work exceeds the number of jobs available at the prevailing wage.

Figure 2.2 shows the overall unemployment in the six decades since the end of World War II (data displayed graphically in Figure 2.2 are contained in a table inside the front cover). The data indicate that through the 1960s, the unemployment rate was usually in the range of 3.5 percent to 5.5 percent, twice going up to around 6.8 percent. In the 1970s, 1980s, and early 1990s, the unemployment rate almost never went below 5.5 percent and went to over 9.5 percent in the early 1980s. The rate was below 5 percent in seven of the eleven years from 1997 through

Figure 4.2

Unemployment Rates for the Civilian Labor Force, 1946–2009 (detailed data in table inside front cover)



2007, before rising to over 9 percent during the latest recession. We will discuss various issues related to unemployment and its measurement in chapter 14.

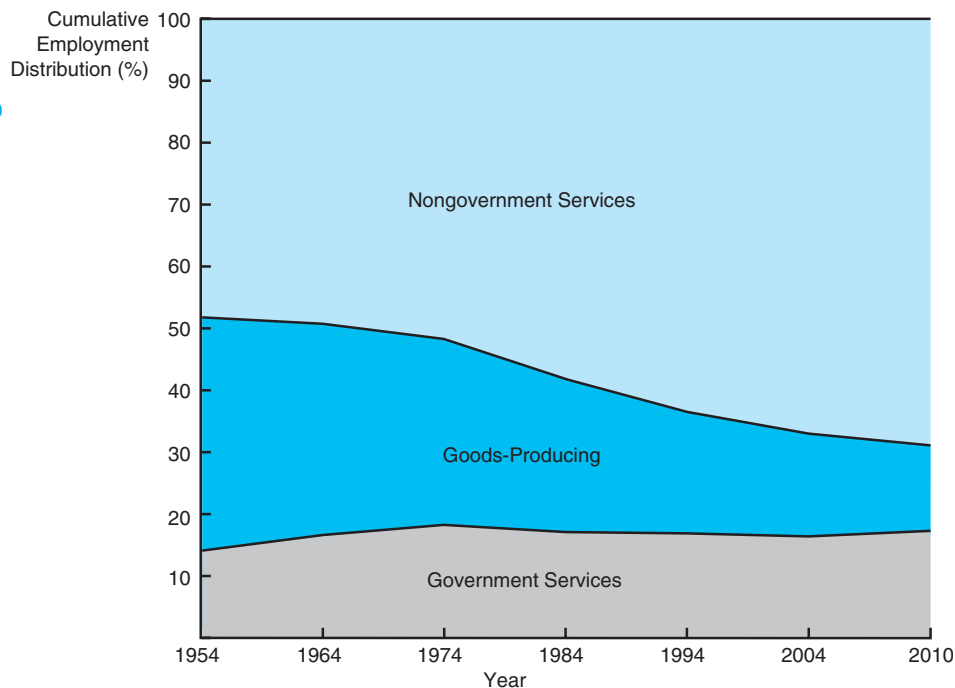
Industries and Occupations: Adapting to Change

As we pointed out earlier, the labor market is the mechanism through which workers and jobs are matched. Over the last half-century, the number of some kinds of jobs has expanded and the number of others has contracted. Both workers and employers have had to adapt to these changes in response to signals provided by the labor market. The labor-market changes occurring in a dynamic economy are sizable; for example, during mid-2007 (before the start of the latest recession), one in every 15 jobs in the United States ended, and about the same fraction was newly created—in just a typical *three-month* period!³

An examination of the industrial distribution of employment from 1954 to 2010 reveals the kinds of changes the labor market has had to facilitate. Figure 2.3,

Figure 4.3

Employment
Distribution by
Major Nonfarm
Sector, 1954–2010
(detailed data in
table inside front
cover)



³U.S. Department of Labor, Bureau of Labor Statistics, "Business Employment Dynamics: Third Quarter 2007," News Release USDL 08-0686 (May 21, 2008), at <http://www.bls.gov>.

which graphs data presented in a table inside the front cover, discloses a major shift: employment in goods-producing industries (largely manufacturing) has fallen as a share of total nonfarm employment, while private-sector services have experienced dramatic growth. Thus, while a smaller share of the American labor force is working in factories, job opportunities with private employers have expanded in wholesale and retail trade, education and health care, professional and business services, leisure and hospitality activities, finance, and information services. Government employment as a share of the total has fluctuated in a relatively narrow range over the period.

The combination of shifts in the industrial distribution of jobs and changes in the production technology within each sector has also required that workers acquire new skills and work in new jobs. Since 1983, for example, the share of American workers in managerial and professional jobs rose from 23 percent to 37 percent, the share in lower-level service jobs rose from 14 percent to almost 18 percent, while the share in administrative-support, sales, and factory jobs fell from 63 percent to 46 percent.⁴

The Earnings of Labor

The actions of buyers and sellers in the labor market serve both to allocate and to set prices for various kinds of labor. From a social perspective, these prices act as signals or incentives in the allocation process—a process that relies primarily on individual and voluntary decisions. From the workers' point of view, the price of labor is important in determining income—and, hence, purchasing power.

Nominal and Real Wages The *wage rate* is the price of labor per working hour.⁵ The *nominal wage* is what workers get paid per hour in current dollars; nominal wages are most useful in comparing the pay of various workers at a given time. *Real wages*, nominal wages divided by some measure of prices, suggest how much can be purchased with workers' nominal wages. For example, if a worker earns \$64 a day and a pair of shoes cost \$32, we could say the worker earns the equivalent of two pairs of shoes a day (real wage = $\$64/\$32 = 2$).

⁴U.S. Department of Labor, Bureau of Labor Statistics, *Employment and Earnings*: 31 (January 1984), Table 20; 57 (January 2010), Table 10.

⁵In this book, we define the hourly wage in the way most workers would if asked to state their "straight-time" wage. It is the money a worker would lose per hour if he or she had an unauthorized absence. When wages are defined in this way, a paid holiday becomes an "employee benefit," as we note in the following, because leisure time is granted while pay continues. Thus, a worker paid \$100 for 25 hours—20 of which are working hours and 5 of which are time off—will be said to earn a wage of \$4 per hour and receive time off worth \$20. An alternative is to define the wage in terms of actual hours worked—or as \$5 per hour in the above example. We prefer our definition, because if the worker seizes an opportunity to work one less hour in a particular week, his or her earnings would fall by \$4, not \$5 (as long as the reduction in hours does not affect the hours of paid holiday or vacation time for which the worker is eligible).

Calculations of real wages are especially useful in comparing the purchasing power of workers' earnings over a period of time when both nominal wages and product prices are changing. For example, suppose we were interested in trying to determine what happened to the real wages of American nonsupervisory workers over the period from 1980 to 2009. We can note from Table 2.2 that the average hourly earnings of these workers in the private sector were \$6.85 in 1980, \$10.20 in 1990, and \$18.60 in 2009; thus, nominal wage rates were clearly rising over this period. However, the prices such workers had to pay for the items they bought were also rising over this period, so a method of accounting for price inflation must be used in calculating real wages.

The most widely used measure for comparing the prices consumers face over several years is the Consumer Price Index (CPI). Generally speaking, this index is derived by determining what a fixed bundle of consumer goods and services (including food, housing, clothing, transportation, medical care, and entertainment) costs each year. The cost of this bundle in the base period is then set to equal 100, and the index numbers for all other years are set proportionately to this base period. For example, if the bundle's average cost over the 1982–1984 period is considered the base (the average value of the index over this period is set to 100), and if the bundle were to cost twice as much in 2009, then the index for 2009 would be set to 200. From the second line in Table 2.2, we can see that with a 1982–1984 base, the CPI was 82.4 in 1980 and 214.5 in 2009—implying that prices had more than doubled ($214.5/82.4 = 2.60$) over that period. Put differently, a dollar in 2009 appears to buy less than half as much as a 1980 dollar.

There are several alternative ways to calculate real wages from the information given in the first two rows of Table 2.2. The most straightforward way is to divide the nominal wage by the CPI for each year and multiply by 100. Doing this converts the nominal wage for each year into 1982–1984 dollars; thus, workers paid \$6.85 in 1980 could have bought \$8.31 worth of goods and services in 1982–1984. Alternatively, we could use the table's information to put average

Table 4.2
Nominal and Real Hourly Earnings, U.S. Nonsupervisory Workers in the Private Sector, 1980–2009

	1980	1990	2009
Average hourly earnings	\$ 6.85	\$10.20	\$18.60
Consumer Price Index (CPI) using 1982–1984 as a base	82.4	130.7	214.5
Average hourly earnings, 1982–1984 dollars (using CPI)	\$ 8.31	\$ 7.80	\$ 8.67
Average hourly earnings, 2009 dollars (using CPI)	\$17.83	\$16.74	\$18.60
Average hourly earnings, 2009 dollars (using CPI inflation less 1 percent per year)	\$13.44	\$13.79	\$18.60

Source: U.S. President, *Economic Report of the President* (Washington, D.C.: U.S. Government Printing Office, 2010), Tables B-47 and B-60.

hourly earnings into 2009 dollars by multiplying each year's nominal wage rate by the price increase between that year and 2009. Because prices rose 2.6 times between 1980 and 2009, \$6.85 in 1980 was equivalent to \$17.83 in 2009.

The CPI Our calculations in Table 2.2 suggest that real wages for American non-supervisory workers were only slightly higher in 2009 than they were in 1980 (and actually fell during the 1980s). A lively debate exists, however, about whether real-wage calculations based on the CPI are accurate indicators of changes in the purchasing power of an hour of work for the ordinary American. The issues are technical and beyond the scope of this text, but they center on two problems associated with using a fixed bundle of goods and services to compare prices from year to year.

One problem is that consumers *change* the bundle of goods and services they actually buy over time, partly in response to changes in prices. If the price of beef rises, for example, consumers may eat more chicken; pricing a fixed bundle may thus understate the purchasing power of current dollars, because it assumes that consumers still purchase the former quantities of beef. For this reason, the bundles used for pricing purposes are updated periodically.

The more difficult issue has to do with the *quality* of goods and services. Suppose that hospital costs rise by 50 percent over a five-year period, but at the same time, new diagnostic equipment and surgical techniques are perfected. Some of the increased price of hospitalization, then, reflects the availability of new services—or quality improvements in previously provided ones—rather than reductions in the purchasing power of a dollar. The problem is that we have not yet found a satisfactory method for feasibly separating the effects of changes in quality.

After considering these problems, some economists believe that the CPI has overstated inflation by as much as one percentage point per year.⁶ While not everyone agrees that inflation is overstated by this much, it is instructive to recalculate real-wage changes by supposing that it is. Inflation, as measured by the CPI, averaged 2.6 percent per year from 1990 to 2009, and in Table 2.2, we therefore estimated that it would take \$16.74 in 2009 to buy what \$10.20 could purchase 19 years earlier. Comparing \$16.74 with what was actually paid in 2009—\$18.60—we would conclude that real wages had risen by 11 percent from 1990 to 2009. If the true decline in purchasing power were instead only 1.6 percent per year during that period, then it would have taken a wage of only \$13.79 in 2009 to match the purchasing power of \$10.20 in 1990. Because workers were actually paid \$18.60 in 2009, assuming that true inflation was one percentage point below that indicated by the CPI, this results in the conclusion that real wages rose by 35 percent (not just 11 percent) over that period! When we make a similar adjustment in

⁶For a review of studies on this topic, see David E. Lebow and Jeremy B. Rudd, "Measurement Error in the Consumer Price Index: Where Do We Stand?" *Journal of Economic Literature* 41 (March 2003): 159–201. These authors place the upward bias in the CPI at between 0.3 percentage points and 1.4 percentage points per year, with the most likely bias being 0.9 percentage points.

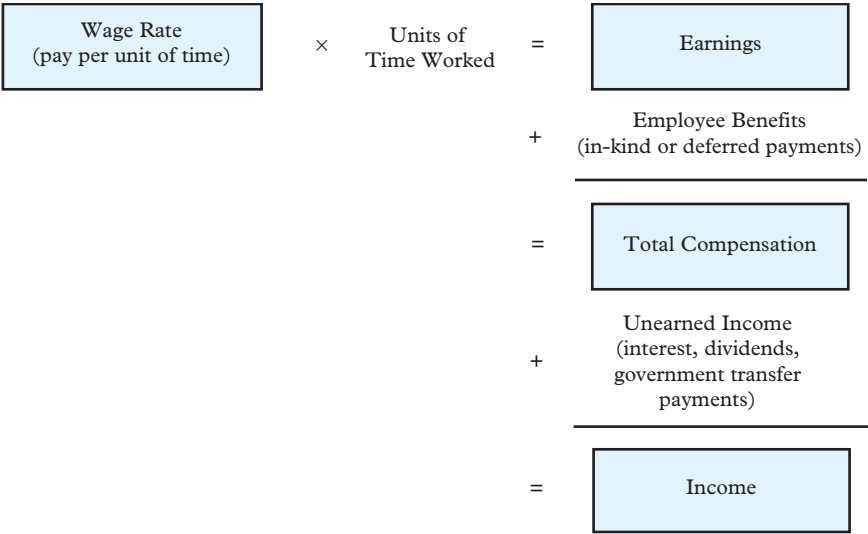
the calculation of real wages for 1980, we estimate that—instead of falling during the 1980s—real wages rose 2.6 percent from 1980 to 1990. Thus, estimated changes in real wage rates are very sensitive to the magnitude of adjustments in the CPI that many economists think should be made.

Wages, Earnings, Compensation, and Income We often apply the term *wages* to payments received by workers who are paid on a salaried basis (monthly, for example) rather than on an hourly basis. The term is used this way merely for convenience and is of no consequence for most purposes. It is important, however, to distinguish among wages, earnings, and income, as we do schematically in Figure 2.4. The term *wages* refers to the payment for a *unit* of time, whereas *earnings* refers to wages multiplied by the number of time units (typically hours) worked. Thus, earnings depend on both wages and the length of time the employee works.

Both wages and earnings are normally defined and measured in terms of direct monetary payments to employees (before taxes for which the employee is liable). *Total compensation*, on the other hand, consists of earnings plus *employee benefits*—benefits that are either payments in kind or deferred. Examples of *payments in kind* are employer-provided health care and health insurance, where the employee receives a service or an insurance policy rather than money. Paid vacation time is also in this category, since employees are given days off instead of cash.

Figure 4.4

Relationship among
Wages, Earnings,
Compensation, and
Income



Deferred payments can take the form of employer-financed retirement benefits, including Social Security taxes, for which employers set aside money now that enables their employees to receive pensions later.

Income—the total command over resources of a person or family during some time period (usually a year)—includes earnings, benefits, and *unearned income*, which includes dividends or interest received on investments and transfer payments received from the government in the form of food stamps, welfare payments, unemployment compensation, and the like.

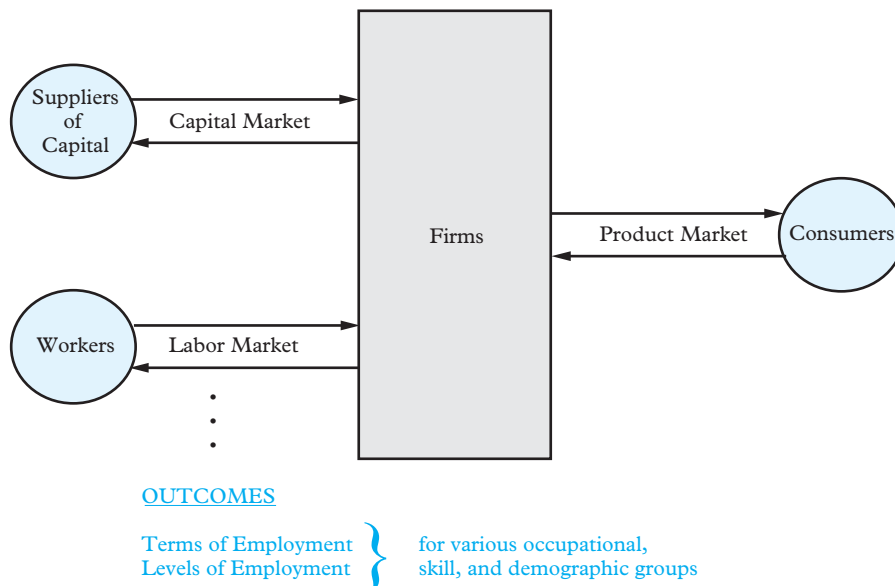
How the Labor Market Works

As shown diagrammatically in Figure 2.5, the labor market is one of three markets in which firms must successfully operate if they are to survive; the other two are the capital market and the product market. The labor and capital markets are the major ones in which firms' inputs are purchased, and the product market is the one in which output is sold. In reality, of course, a firm may deal in many different labor, capital, or product markets simultaneously.

Study of the labor market begins and ends with an analysis of the demand for and supply of labor. On the demand side of the labor market are employers, whose decisions about the hiring of labor are influenced by conditions in all three markets. On the supply side of the labor market are workers and potential

Figure 4.5

The Markets in Which Firms Must Operate



workers, whose decisions about where (and whether) to work must take into account their other options for how to spend time.

It is useful to remember that the major labor market outcomes are related to (a) the *terms of employment* (wages, compensation levels, working conditions) and (b) the *levels of employment*. In analyzing both these outcomes, one must usually differentiate among the various occupational, skill, or demographic groups that make up the overall labor market. Any labor market outcome is always affected, to one degree or another, by the forces of both demand and supply. To paraphrase economist Alfred Marshall, it takes both demand and supply to determine economic outcomes, just as it takes both blades of a scissors to cut cloth.

In this chapter, we present the basic outlines and broadest implications of the simplest economic model of the labor market. In later chapters, we shall add some complexities to this basic model and explain assumptions and implications more fully. However, the simple model of demand and supply presented here offers some insights into labor market behavior that can be very useful in the formulation of social policy. Every piece of analysis in this text is an extension or modification of the basic model presented in this chapter.

The Demand for Labor

Firms combine various factors of production—mainly capital and labor—to produce goods or services that are sold in a product market. Their total output and the way in which they combine labor and capital depend on three forces: product demand, the amount of labor and capital they can acquire at given prices, and the choice of technologies available to them. When we study the demand for labor, we are interested in finding out how the number of workers employed by a firm or set of firms is affected by changes in one or more of these three forces. To simplify the discussion, we shall study one change at a time while holding other forces constant.

Wage Changes How does the number of employees (or total labor hours) demanded vary when wages change? Suppose, for example, that we could vary the wages facing a certain industry over a long period of time but keep the technology available, the conditions under which capital is supplied, and the relationship between product price and product demand remain unchanged. What would happen to the quantity of labor demanded if the wage rate were *increased*?

First, higher wages imply higher costs and, usually, higher product prices. Because consumers respond to higher prices by buying less, employers would tend to reduce their levels of output and employment (other things being equal). This decline in employment is called a *scale effect*—the effect on desired employment of a smaller scale of production.

Second, as wages increase (assuming the price of capital does not change, at least initially), employers have incentives to cut costs by adopting a technology that relies more on capital and less on labor. Desired employment would fall because of a shift toward a more *capital-intensive* mode of production. This second

Table 4.3**Labor Demand Schedule for a Hypothetical Industry**

Wage Rate (\$)	Desired Employment Level
3.00	250
4.00	190
5.00	160
6.00	130
7.00	100
8.00	70

Note: Employment levels can be measured in number of employees or number of labor hours demanded. We have chosen here to use number of employees.

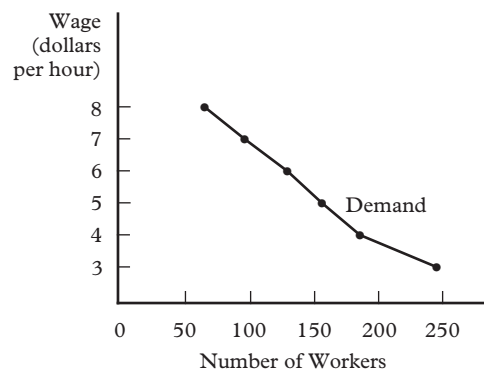
effect is termed a *substitution effect*, because as wages rise, capital is *substituted* for labor in the production process.

The effects of various wages on employment levels might be summarized in a table showing the labor demanded at each wage level. Table 2.3 illustrates such a *demand schedule*. The relationship between wages and employment tabulated in Table 2.3 could be graphed as a *demand curve*. Figure 2.6 shows the demand curve generated by the data in Table 2.3. Note that the curve has a negative slope, indicating that as wages rise, less labor is demanded. (Note also that we follow convention in economics by placing the wage rate on the *vertical* axis despite its being an *independent* variable in the context of labor demand by a firm.) A demand curve for labor tells us how the desired level of employment, measured in either labor hours or number of employees, varies with changes in the price of labor when the other forces affecting demand are held constant.

Changes in Other Forces Affecting Demand What happens to labor demand when one of the forces other than the wage rate changes?

Figure 4.6

Labor Demand Curve (based on data in Table 2.3)



First, suppose that *demand for the product* of a particular industry were to increase, so that at any output price, more of the goods or services in question could be sold. Suppose in this case that technology and the conditions under which capital and labor are made available to the industry do not change. Output levels would clearly rise as firms in the industry sought to maximize profits, and this *scale* (or *output*) *effect* would increase the demand for labor at any given wage rate. (As long as the relative prices of capital and labor remain unchanged, there is no *substitution effect*.)

How would this change in the demand for labor be illustrated using a demand curve? Since the technology available and the conditions under which capital and labor are supplied have remained constant, this change in product demand would increase the labor desired at any wage level that might prevail. In other words, the entire labor demand curve *shifts* to the right. This rightward shift, shown as a movement from D to D' in Figure 2.7, indicates that at every possible wage rate, the number of workers demanded has increased.

Second, consider what would happen if the product demand schedule, technology, and labor supply conditions were to remain unchanged, but *the supply of capital* changed so that capital prices fell to 50 percent of their prior level. How would this change affect the demand for labor?

Our method of analyzing the effects on labor demand of a change in the price of *another* productive input is familiar: we must consider the scale and substitution effects. First, when capital prices decline, the costs of producing tend to decline. Reduced costs stimulate increases in production, and these increases tend to raise the level of desired employment at any given wage. The scale effect of a fall in capital prices thus tends to increase the demand for labor at each wage level.

The second effect of a fall in capital prices would be a substitution effect, whereby firms adopt more capital-intensive technologies in response to cheaper capital. Such firms would substitute capital for labor and would use less labor to produce a given amount of output than before. With less labor being desired at each wage rate and output level, the labor demand curve tends to shift to the left.

Figure 4.7

Shift in Demand for Labor Due to Increase in Product Demand

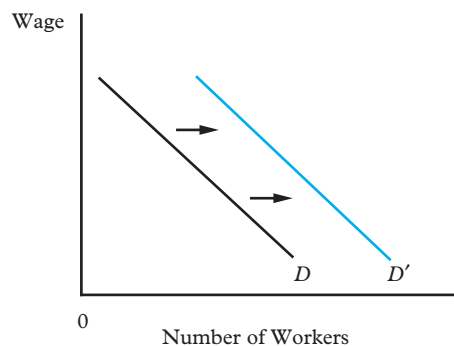
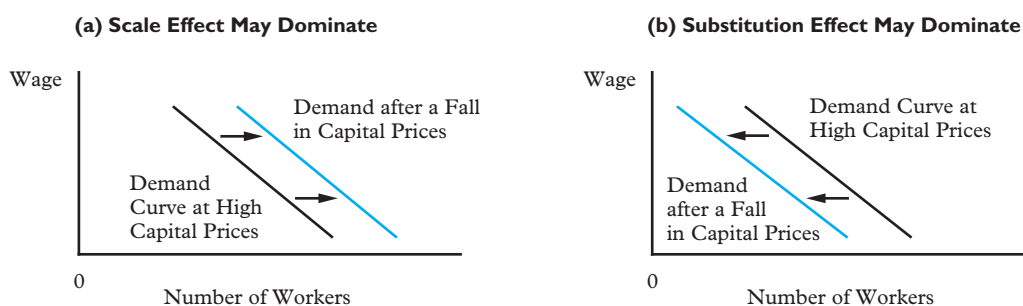


Figure 4.8

Possible Shifts in Demand for Labor Due to Fall in Capital Prices



A fall in capital prices, then, generates *two opposite effects* on the demand for labor. The scale effect will push the labor demand curve rightward, while the substitution effect will push it to the left. As emphasized by Figure 2.8, either effect could dominate. Thus, economic theory does not yield a clear-cut prediction about how a fall in capital prices will affect the demand for labor. (A *rise* in capital prices would generate the same overall ambiguity of effect on the demand for labor, with the scale effect pushing the labor demand curve leftward and the substitution effect pushing it to the right.)

The hypothesized changes in product demand and capital supply just discussed have tended to *shift* the demand curve for labor. It is important to distinguish between a *shift* in a demand curve and *movement along* a curve. A labor demand curve graphically shows the *labor desired* as a function of the *wage rate*. When the *wage* changes and other forces are held unchanged, one *moves along* the curve. However, when one of the *other forces* changes, the labor demand curve *shifts*. Unlike wages, these forces are not directly shown when the demand curve for labor is drawn. Thus, when *they* change, a different relationship between wages and desired employment prevails, and this shows up as a shift of the demand curve.

Market, Industry, and Firm Demand The demand for labor can be analyzed on three levels:

1. To analyze the demand for labor *by a particular firm*, we would examine how an increase in the wage of machinists, say, would affect their employment by a particular aircraft manufacturer.
2. To analyze the effects of this wage increase on the employment of machinists *in the entire aircraft industry*, we would utilize an industry demand curve.
3. Finally, to see how the wage increase would affect the *entire labor market* for machinists in all industries in which they are used, we would use a market demand curve.

We shall see in chapters 3 and 4 that firm, industry, and market labor demand curves vary in *shape* to some extent because *scale* and *substitution effects* have different strengths at each level. However, it is important to remember that the scale and substitution effects of a wage change work in the same direction at each level, so that firm, industry, and market demand curves *all slope downward*.

Long Run versus Short Run We can also distinguish between *long-run* and *short-run* labor demand curves. Over very short periods of time, employers find it difficult to substitute capital for labor (or vice versa), and customers may not change their product demand very much in response to a price increase. It takes *time* to fully adjust consumption and production behavior. Over longer periods of time, of course, responses to changes in wages or other forces affecting the demand for labor are larger and more complete.

The Supply of Labor

Having looked at a simple model of behavior on the buyer (or demand) side of the labor market, we now turn to the seller (or supply) side of the market. For the purposes of this chapter, we shall assume that workers have already decided to work and that the question facing them is what occupation and what employer to choose.

Market Supply To first consider the supply of labor to the entire market (as opposed to the supply to a particular firm), suppose that the market we are considering is the one for legal assistants (or “paralegals”). How will supply respond to changes in the wages paralegals might receive?

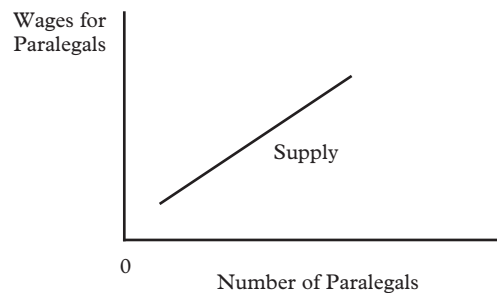
If the salaries and wages in *other* occupations are held constant and the wages of paralegals rise, we would expect to find more people wanting to become paralegals. For example, suppose that each of 100 people in a high school graduating class has the option of becoming an insurance agent or a paralegal. Some of these 100 people will prefer to be insurance agents even if paralegals are better paid, because they like the challenge and sociability of selling. Some would want to be paralegals even if the pay were comparatively poor, because they hate the pressures of selling. Many, however, could see themselves doing either job; for them, the compensation in each occupation would be a major factor in their decision.

Thus, the supply of labor to a particular market is positively related to the wage rate prevailing in that market, holding other wages constant. That is, if the wages of insurance agents are held constant and the paralegal wage rises, more people will want to become paralegals because of the relative improvement in compensation (as shown graphically in Figure 2.9).

As with demand curves, each supply curve is drawn holding other prices and wages constant. If one or more of these other prices or wages were to change, it would cause the supply curve to *shift*. As the salaries of insurance agents *rise*, some people will change their minds about becoming paralegals and choose to become

Figure 4.9

Market Supply Curve for Paralegals



insurance agents. In graphical terms (see Figure 2.10), increases in the salaries of insurance agents would cause the supply curve of paralegals to shift to the left.

Supply to Firms Having decided to become a paralegal, an individual would then have to decide which offer of employment to accept. If all employers were offering paralegal jobs that were more or less alike, the choice would be based entirely on compensation. Any firm unwise enough to attempt paying a wage below what others are paying would find it could not attract any employees (or at least none of the caliber it wanted). Conversely, no firm would be foolish enough to pay more than the going wage, because it would be paying more than it would have to pay to attract a suitable number and quality of employees. Supply curves to a *firm*, then, are *horizontal*, as shown in Figure 2.11, indicating that at the going wage, a firm could get all the paralegals it needs. If the paralegal wage paid by others in the market is W_0 , then the firm's labor supply curve is S_0 ; if the wage falls to W_1 , the firm's labor supply curve becomes S_1 .

The difference in slope between the market supply curve and the supply curve to a firm is directly related to the type of choice facing workers. In deciding whether to *enter* the paralegal labor market, workers must weigh both the compensation *and* the job requirements of alternative options (such as being an

Figure 4.10

Shift in Market Supply Curve for Paralegals as Salaries of Insurance Agents Rise

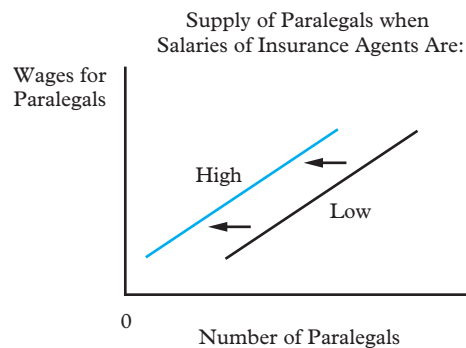
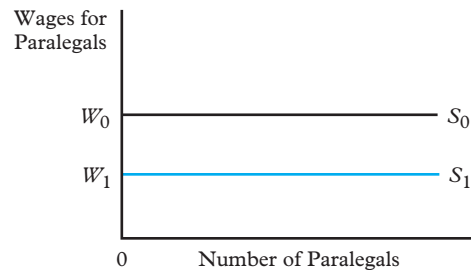


Figure 4.11

Supply of Paralegals to a Firm at Alternative Market Wages



insurance agent). If wages for paralegals were to fall, not everyone would withdraw from that market, because the jobs of insurance agent and paralegal are not perfect substitutes. Some people would remain paralegals after a wage decline because they dislike the job requirements of insurance agents.

Once the decision to become a paralegal had been made, however, the choice of *which employer* to work for would be a choice among alternatives in which the job requirements were nearly the *same*. Thus, the choice would have to be made on the basis of compensation alone. If a firm were to lower its wage offers below those of other firms, it would lose all its applicants. The horizontal supply curve is, therefore, a reflection of supply decisions made among alternatives that are perfect substitutes for each other.

We have argued that firms wishing to hire paralegals must pay the going wage or lose all applicants. While this may seem unrealistic, it is not a bad proposition with which to start our analysis. If a firm offers jobs *comparable* to those offered by other firms but at a lower level of pay, it might be able to attract a few applicants of the quality it desires because a few people will be unaware of compensation elsewhere. Over time, however, knowledge of the firm's poor pay would become more widespread, and the firm would have to rely solely on less-qualified people to fill its jobs. It could secure quality employees at below-average pay only if it offered *noncomparable* jobs (more pleasant working conditions, longer paid vacations, and so forth). This factor in labor supply will be discussed in chapter 8. For now, we will assume that individual firms, like individual workers, are *wage takers*; that is, the wages they pay to their workers must be pretty close to the going wage if they face competition in the labor market. Neither individual workers nor firms can set a wage much different from the going wage and still hope to transact. (Exceptions to this elementary proposition will be analyzed in chapter 5.)

The Determination of the Wage

The wage that prevails in a particular labor market is heavily influenced by labor supply and demand, regardless of whether the market involves a labor union or other nonmarket forces. In this section, we analyze how the interplay of supply and demand in the labor market affects wages.

The Market-Clearing Wage Recall that the market demand curve indicates how many workers employers would want at each wage rate, holding capital prices and the product demand schedule constant. The market supply curve indicates how many workers would enter the market at each wage level, holding the wages in other occupations constant. These curves can be placed on the same graph to reveal some interesting information, as shown in Figure 2.12.

For example, suppose the market wage were set at W_1 . At this low wage, Figure 2.12 indicates that demand *exceeds* supply. Employers will be competing for the few workers in the market, and a shortage of workers would exist. The desire of firms to attract more employees would lead them to increase their wage offers, thus driving up the overall level of wage offers in the market. As wages rose, two things would happen. First, more workers would choose to enter the market and look for jobs (a movement along the supply curve); second, increasing wages would induce employers to seek fewer workers (a movement along the demand curve).

If wages were to rise to W_2 , supply would exceed demand. Employers would desire fewer workers than the number available, and not all those desiring employment would be able to find jobs, resulting in a surplus of workers. Employers would have long lines of eager applicants for any opening and would find that they could fill their openings with qualified applicants even if they offered lower wages. Furthermore, if they could pay lower wages, they would want to hire more employees. Some employees would be more than happy to accept lower wages if they could just find a job. Others would leave the market and look for work elsewhere as wages fell. Thus, supply and demand would become more equal as wages fell from the level of W_2 .

The wage rate at which demand equals supply is the *market-clearing* wage. At W_e in Figure 2.12, employers can fill the number of openings they have, and all employees who want jobs in this market can find them. At W_e there is no surplus and no shortage. All parties are satisfied, and no forces exist that would alter the wage. The market is in equilibrium in the sense that the wage will remain at W_e .

Figure 4.12

Market Demand and Supply

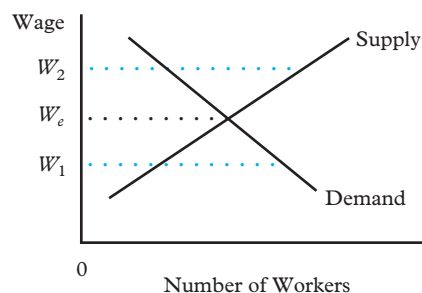
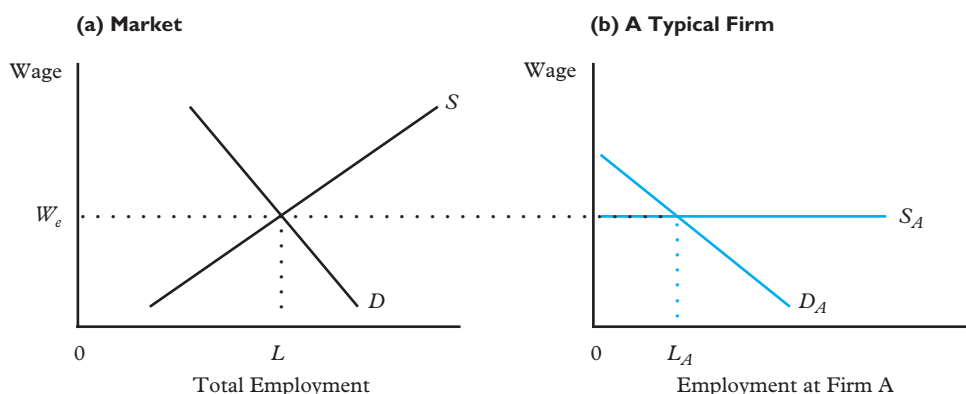


Figure 4.13

Demand and Supply at the “Market” and “Firm” Levels

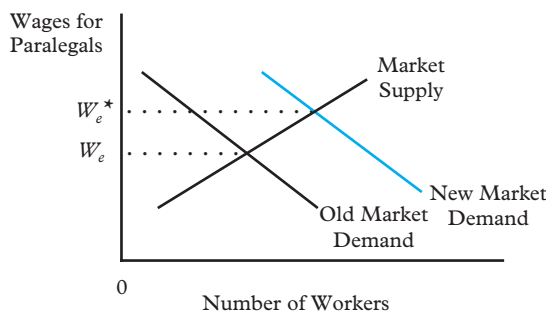


The market-clearing wage, W_e , thus becomes the *going wage* that individual employers and employees must face. In other words, wage rates are determined by the market and “announced” to individual market participants. Figure 2.13 graphically depicts *market* supply and demand in panel (a), along with the supply and demand curves for a typical *firm* (firm A) in that market in panel (b). All firms in the market pay a wage of W_e , and total employment of L equals the sum of employment in each firm.

Disturbing the Equilibrium What could happen to change the market-clearing wage once it has been reached? Changes could arise from shifts in either the demand or the supply curve. Suppose, for example, that the increase in paper-work accompanying greater government regulation of industry caused firms to demand more paralegal help (at any given wage rate) than before. Graphically, as in Figure 2.14, this greater demand would be represented as a rightward shift of

Figure 4.14

New Labor Market Equilibrium after Demand Shifts Right



the labor demand curve. If W_e were to persist, there would be a labor shortage in the paralegal market (because demand would exceed supply). This shortage would induce employers to improve their wage offers. Eventually, the paralegal wage would be driven up to W_e^* . Notice that in this case, the equilibrium level of *employment* will also rise.

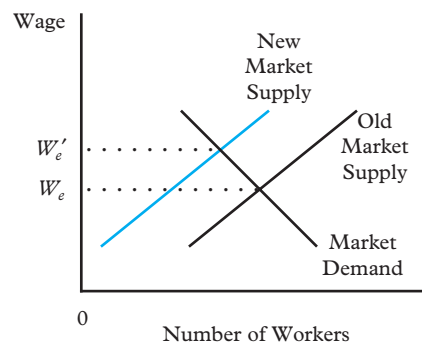
The market wage can also increase if the labor supply curve shifts to the left. As shown in Figure 2.15, such a shift creates a labor shortage at the old equilibrium wage of W_e , and as employers scramble to fill their job openings, the market wage is bid up to W_e' . In the case of a leftward-shifting labor supply curve, however, the increased market wage is accompanied by a decrease in the equilibrium level of employment. (See Example 2.1 for an analysis of the labor market effects of the leftward shift in labor supply accompanying the Black Death in 1348–1351.)

If a leftward shift in labor supply is accompanied by a rightward shift in labor demand, the market wage can rise dramatically. Such a condition occurred in Egypt during the early 1970s. Lured by wages over six times higher in Saudi Arabia and other oil-rich Arab countries, roughly half of Egypt's construction workers left the country just as a residential building boom in Egypt got under way. The combination of a leftward-shifting labor supply curve and a rightward-shifting labor demand curve drove the real wages of Egyptian construction workers up by over 100 percent in just five years!⁷ (This notable wage increase was accompanied by a net employment *increase* in Egypt's construction industry. The student will be asked in the first review question on page 55 to analyze these events graphically.)

A fall in the market-clearing wage rate would occur if there were increased supply or reduced demand. An increase in supply would be represented by a rightward shift of the supply curve, as more people entered the market at each

Figure 4.15

New Labor Market Equilibrium after Supply Shifts Left



⁷Bent Hansen and Samir Radwan, *Employment Opportunities and Equity in Egypt* (Geneva: International Labour Office, 1982): 74.

EXAMPLE 4.1**The Black Death and the Wages of Labor**

An example of what happens to wages when the supply of labor suddenly shifts occurred when plague—the Black Death—struck England (among other European countries) in 1348–1351. Estimates vary, but it is generally agreed that plague killed between 17 percent and 40 percent of the English population in that short period of time. This shocking loss of life had the immediate effect of raising the wages of laborers. As the supply curve shifted to the left, a shortage of workers was created at the old wage levels, and competition among employers for the surviving workers drove the wage level dramatically upward.

Reliable figures are hard to come by, but many believe wages rose by 50–100 percent over the four-year period. A thresher, for example, earning 2½ pence per day in 1348 earned 4½ pence in 1350, and mowers receiving 5 pence per acre in 1348 were receiving 9 pence in 1350. Whether the overall rise in wages was this large or not, there was clearly a labor shortage and an unprecedented increase in wages. A royal proclamation commanding landlords to share their scarce workers with neighbors and threatening workers with imprisonment if they refused work at the pre-plague wage was issued to deal with this shortage, but it was ignored. The shortage was too severe

and market forces were simply too strong for the rise in wages to be thwarted.

The discerning student might wonder at this point about the demand curve for labor. Did it not also shift to the left as the population—and the number of consumers—declined? It did, but this leftward shift was not as pronounced as the leftward shift in supply. While there were fewer customers for labor's output, the customers who remained consumed greater amounts of goods and services per capita than before. The money, gold and silver, and durable goods that had existed prior to 1348 were divided among many fewer people by 1350, and this rise in per capita wealth was associated with a widespread and dramatic increase in the level of consumption, especially of luxury goods. Thus, the leftward shift in labor demand was dominated by the leftward shift in supply, and the predictable result was a large increase in wages.

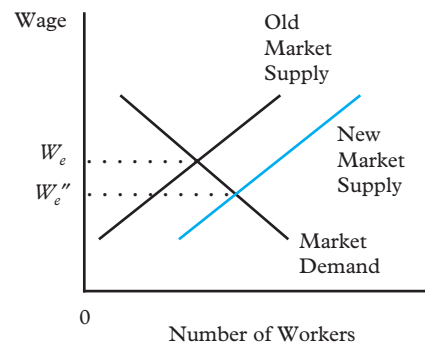
Data from: Harry A. Miskimin, *The Economy of Early Renaissance Europe, 1300–1460* (Englewood Cliffs, N.J.: Prentice-Hall, 1969); George M. Modlin and Frank T. deVyver, *Development of Economic Society* (Boston: D.C. Heath, 1946); Douglass C. North and Robert Paul Thomas, *The Rise of the Western World* (Cambridge: Cambridge University Press, 1973); Philip Ziegler, *The Black Death* (New York: Harper and Row, 1969).

wage (see Figure 2.16). This rightward shift would cause a surplus to exist at the old equilibrium wage (W_e) and lead to behavior that reduced the wage to W_e'' in Figure 2.16. Note that the equilibrium employment level has increased. A decrease (leftward shift) in labor demand would also cause a decrease in the market-clearing wage, although such a shift would be accompanied by a *fall* in employment.

Disequilibrium and Nonmarket Influences That a market-clearing wage exists in theory does not imply that it is reached—or reached quickly—in practice. Because labor services cannot be separated from the worker, and because labor income is by far the most important source of spending power for ordinary people, the labor market is subject to forces that impede the adjustment of both wages and employment to changes in supply or demand. Some of these barriers to adjustment are themselves the result of economic forces that will be discussed later in the text. For example, changing jobs often requires an employee to invest in new skills (see

Figure 4.16

New Labor Market Equilibrium after Supply Shifts Right



chapter 9) or bear costs of moving (chapter 10). On the employer side of the market, hiring workers can involve an initial investment in search and training (chapter 5), while firing them or cutting their wages can be perceived as unfair and therefore have consequences for the productivity of those who remain (chapter 11).

Other barriers to adjustment are rooted in *nonmarket* forces: laws, customs, or institutions constraining the choices of individuals and firms. Although forces keeping wages *below* their market-clearing levels are not unknown, nonmarket forces usually serve to keep wages *above* market levels. Minimum wage laws (discussed in chapter 4) and unions (chapter 13) are examples of influences explicitly designed to raise wages beyond those dictated by the market. Likewise, if there is a widespread belief that cutting wages is unfair, laws or customs may arise that prevent wages from falling in markets experiencing leftward shifts in demand or rightward shifts in supply.

It is commonly believed that labor markets adjust more quickly when market forces are calling for wages to rise as opposed to pressuring them to fall. If this is so, then those markets observed to be in disequilibrium for long periods will tend to be ones with above-market wages. The existence of above-market wages implies that the supply of labor exceeds the number of jobs being offered (refer to the relative demand and supply at wage W_2 in Figure 2.12); therefore, if enough markets are experiencing above-market wages the result will be widespread *unemployment*. In fact, as we will see in the section International Differences in Unemployment, these differences can sometimes be used to identify where market forces are most constrained by nonmarket influences.

Applications of the Theory

Although this simple model of how a labor market functions will be refined and elaborated upon in the following chapters, it can explain many important phenomena, including the issues of when workers are overpaid or underpaid and what explains international differences in unemployment.

Who Is Underpaid and Who Is Overpaid?

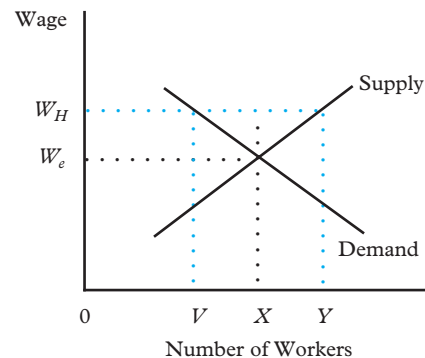
We pointed out in chapter 1 that a fundamental value of normative economics is that, as a society, we should strive to complete all those transactions that are mutually beneficial. Another way of stating this value is to say that we must strive to use our scarce resources as effectively as possible, which implies that output should be produced in the least-costly manner so that the most can be obtained from such resources. This goal, combined with the labor market model outlined in this chapter, suggests how we can define what it means to be overpaid.

Above-Market Wages We shall define workers as *overpaid* if their wages are higher than the market-clearing wage for their job. Because a labor surplus exists for jobs that are overpaid, a wage above market has two implications (see Figure 2.17). First, employers are paying more than necessary to produce their output (they pay W_H instead of W_e); they could cut wages and still find enough qualified workers for their job openings. In fact, if they did cut wages, they could expand output and make their product cheaper and more accessible to consumers. Second, more workers want jobs than can find them (Y workers want jobs, but only V openings are available). If wages were reduced a little, more of these disappointed workers could find work. A wage above market thus causes consumer prices to be higher and output to be smaller than is possible, and it creates a situation in which not all workers who want the jobs in question can get them.

An interesting example of above-market wages was seen in Houston's labor market in 1988. Bus cleaners working for the Houston Metropolitan Transit Authority received \$10.08 per hour, or 70 percent more than the \$5.94 received by cleaners working for private bus companies in Houston. One

Figure 4.17

Effects of an Above-Market Wage



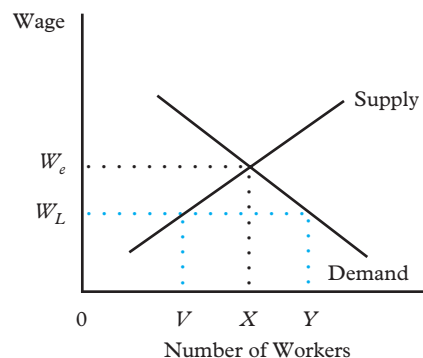
(predictable) result of this overpayment is that the quit rate among Houston's Transit Authority cleaners was only *one-seventh* as great as the average for cleaners nationwide.⁸

To better understand the social losses attendant on overpayment, let us return to the principles of normative economics. Can reducing overpayment create a situation in which the gainers gain more than the losers lose? Suppose in the case of Houston's Transit Authority cleaners that *only* the wage of *newly hired* cleaners was reduced—to \$6.40, say. Current cleaners thus would not lose, but many others who were working elsewhere at \$5.94 would jump at the chance to earn a higher wage. Taxpayers, realizing that transit services could now be expanded at lower cost than before, would increase their demand for such services, thus creating jobs for these additional workers. Some workers would gain, while no one lost—and social well-being would clearly be enhanced.⁹ The wage reduction, in short, would be *Pareto-improving* (see chapter 1).

Below-Market Wages Employees can be defined as *underpaid* if their wage is below market-clearing levels. At below-market wages, employers have difficulty finding workers to meet the demands of consumers, and a labor shortage thus exists. They also have trouble keeping the workers they do find. If wages were increased, output would rise and more workers would be attracted to the market. Thus, an increase would benefit the people in society in *both* their consumer and their worker roles. Figure 2.18 shows how a wage increase from W_L to W_e would increase employment from V to X (at the same time wages were rising).

Figure 4.18

Effects of a Below-Equilibrium Wage



⁸William J. Moore and Robert J. Newman, "Government Wage Differentials in a Municipal Labor Market: The Case of Houston Metropolitan Transit Workers," *Industrial and Labor Relations Review* 45 (October 1991): 145–153.

⁹If the workers who switched jobs were getting paid approximately what they were worth to their former employers, these employers would lose \$5.94 in output but save \$5.94 in costs—and their welfare would thus not be affected. The presumption that employees are paid what they are worth to the employer is discussed at length in chapter 3.

EXAMPLE 4.2**Forced Labor in Colonial Mozambique**

Two ways to address a labor shortage are to raise wages by enough to attract workers voluntarily into the job or to force workers (by drafting them) into the job. While forced labor may seem to be the cheaper alternative, the resentful workforce that accompanies compulsion carries with it opportunity costs that outweigh the wage savings. An early example can be found in colonial Mozambique.

In the late nineteenth century, Mozambique—which was ruled by Portugal—was divided into several large estates for administrative purposes. The local estate holders owed the colonial administration rent and taxes, but they had the right to collect (and keep) a “head tax” of 800 *reis* per year from each African living within their boundaries. The low wages and harsh working conditions on sugar plantations created a labor shortage on many estates, and in 1880, many estate holders decided to collect the head tax by *forcing Africans to work on their plantation (without pay) for two weeks per year*.

The implied wage rate for these two weeks was 400 *reis* per week, which compares to wages of 500–750 *reis* per week in areas where plantation

labor was recruited through voluntary means. Not surprisingly, estate holders who used forced labor had to contend with a very dissatisfied, resentful group of workers. Their workforce turned over every two weeks, motivation was a problem (causing them to resort to beatings), and they had to employ private police to track down runaways who were seeking to avoid the low implicit pay and harsh methods of motivation.

In 1894, the Mozambique Sugar Company abandoned the use of forced labor, which it found to have very high opportunity costs, and raised wages by enough that workers voluntarily returned to their estates. In essence, then, the estate holders in Mozambique came to the conclusion that it was more profitable to pay the wages they needed to attract a voluntary workforce than to make use of forced labor.

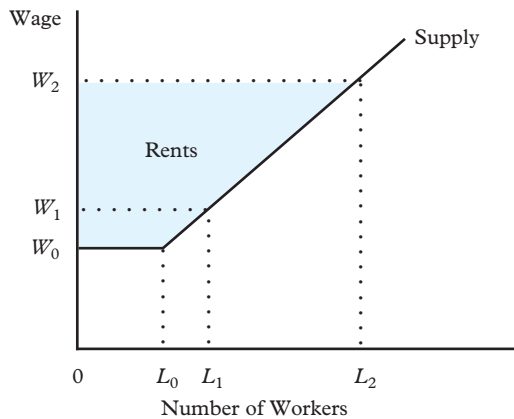
Source: Leroy Vail and Landeg White, *Capitalism and Colonialism in Mozambique: A Study of the Quelimane District* (Minneapolis: University of Minnesota Press, 1980): 77, 120–25, 134.

Wages in the U.S. Army illustrate how the market adjusts to below-market wages. Prior to 1973, when the military draft was eliminated, the government could pursue a policy of paying below-market wages to military recruits, because the resultant gap between supply and demand could be filled by conscription. Not surprisingly, when comparing wages in the late 1970s with those in the last decade of the military draft, we find that the average military cash wages paid to enlisted personnel rose 19 percent more than those of comparable civilian workers. (See Example 2.2 for other labor market effects of relying on forced labor.)

Economic Rents The concepts of underpayment and overpayment have to do with the *social* issue of producing desired goods and services in the least-costly way; therefore, we compared wages paid with the *market-clearing wage*. At the level of *individuals*, however, it is often useful to compare the wage received in a job with one’s *reservation wage*, the wage below which the worker would refuse (or quit) the job in question. The amount by which one’s wage exceeds one’s reservation wage in a particular job is the amount of his or her *economic rent*.

Figure 4.19

Labor Supply to the Military: Different Preferences Imply Different “Rents”



Consider the labor supply curve to, say, the military. As shown in Figure 2.19, if the military is to hire L_1 people, it must pay W_1 in wages. These relatively low wages will attract to the military those who most enjoy the military culture and are least averse to the risks of combat. If the military is to be somewhat larger and to employ L_2 people, then it must pay a wage of W_2 . This higher wage is required to attract those who would have found a military career unattractive at the lower wage. If W_2 turns out to be the wage that equates supply and demand, and if the military pays that wage, everyone who would have joined up for less would be receiving an economic rent!

Put differently, the supply curve to an occupation or industry is a schedule of reservation wages that indicates the labor forthcoming at each wage level. The difference between the wage actually paid and workers' reservation wages—the shaded area in Figure 2.19—is the amount of the rent. Since each worker potentially has a different reservation wage, rents may well differ for each worker in the market. In Figure 2.19, the greatest rents are received by those L_0 individuals who would have joined the military even if the wage were only W_0 . They collect an economic rent of $W_2 - W_0$.

Why don't employers reduce the wage of each employee down to his or her reservation level? While capturing employee rents would seem to be lucrative, since by definition it could be done without the workers' quitting, attempting to do so would create resentment, and such a policy would be extremely costly, if not impossible, to implement. Employers do not know the true reservation wages of each employee or applicant, and finding it would involve experiments in which the wage offers to each worker either started high and were cut or started low and were raised. This would be costly, and if workers realized the firm was *experimenting*, they would attempt to disguise their true reservation wages and adopt the strategic behavior associated with bargaining (bluffing, for example). Therefore, firms usually pay according to the job, one's level of experience or

EMPIRICAL STUDY

PAY LEVELS AND THE SUPPLY OF MILITARY OFFICERS: OBTAINING SAMPLE VARIATION FROM CROSS-SECTION DATA

Economic theory predicts that the supply to a particular occupation is expected to increase when the pay for that occupation increases or when the pay in alternative occupations falls. In the late 1960s, the U.S. government was considering a policy change that eventually resulted in the elimination of the military draft, and it needed to estimate how much military pay would have to rise—relative to civilian pay—to attract the needed number of officers and enlisted personnel without the presence of a draft. Estimating the labor supply curve of, say, officers depends on whether we can obtain an appropriate data set.

Any study of how (independent) variable X affects (dependent) variable Y requires that the researcher have access to a data set in which both X and Y show considerable *variation*. Put differently, scientific research into cause and effect requires that we observe how *different* causes produce *different* effects! Researchers who are able to conduct laboratory experiments expose their subjects to different “treatments” and then look for differences in outcomes. Economists are rarely able to conduct experiments, so they must look for data sets in which X and Y naturally differ across the observations in a sample. If the ratio of military pay to civilian pay is our independent variable (X), and the number of people who decide to join the military as officers is our dependent variable (Y), how can

we generate a sample in which both variables display enough variation to estimate a relationship?

One way is to use data over a period of 20–30 years (“time series” data), with each year’s relative wage and number of new officers representing one observation in the sample. The problem with a time series is that samples are necessarily small (there are not that many years for which we have good data). Behavior can also be affected by all kinds of changing conditions or preferences over time (for example, wars, new occupations both in and out of the military, changing attitudes of the labor force toward risk), so that with time series data, we also need to control for these time-related changes to be confident we have isolated the effects of *pay* on labor supply decisions.

Another way to study the effects of relative pay on labor supply is to use “cross-section” data, which involves collecting observations on pay and labor supply for different people at one point in time. This usually allows for a much larger data set, but it requires that those in the data set be operating in sufficiently different environments that X and Y will actually vary. Within any year, for example, military pay for entry-level officers is the same for everyone, so we can use cross-section data to study military supply decisions only if the *civilian wages* facing sample members

can be accurately measured and turn out to vary significantly.

One study done in the late 1960s analyzed enrollment data from 82 Reserve Officer Training Corps (ROTC) programs offered by universities in 1963. The supply variable (Y) in this study was measured as the percentage of men at each of the 82 universities enrolled in an Army, Navy, or Air Force ROTC program (the military was virtually all male at that time). Because military pay facing ROTC graduates at each of the 82 institutions was the same, differences in civilian pay opportunities for recent graduates represented the only pay variable that could be used. It turned out that the average earnings of recent male college graduates from each of the 82 universities were both available and varied enough across the universities to be useful; thus, the variable measuring pay (X) was the average earnings in 1963 of men who graduated from each of the universities in 1958.

Theory leads us to expect that the higher civilian pay was for the graduates

of a university, the lower would be its ROTC enrollments. The results estimated that there was indeed a negative and statistically significant relationship between civilian pay and ROTC enrollments.^a The size of the estimated relationship suggested that where civilian pay was 10 percent higher, ROTC enrollments were 20 percent lower. This finding implies that if military pay were to have risen by 10 percent, holding civilian pay constant, ROTC enrollments would have risen by 20 percent. Clearly, ROTC enrollments were very responsive to civilian salaries!

^aOther independent variables were added to the estimating equation to account for the fact that the universities sampled offered different mixes of Army, Navy, and Air Force ROTC programs. Furthermore, because students in the South may have had a greater preference for military service at any pay level, the list of independent variables also included a variable indicating if the university was located in the South.

Source: Stuart H. Altman and Alan E. Fechter, "The Supply of Military Personnel in the Absence of a Draft," *American Economic Review* 57 (May 1967): 19–31.

longevity with the employer, and considerations of merit—but not according to preferences.

International Differences in Unemployment

We noted earlier that labor markets are often influenced by nonmarket forces that keep wages above market-clearing levels. Because these nonmarket forces generally take the form of laws, government programs, customs, or institutions (labor unions, for example), their strength typically varies across countries. Can we form some conclusions about the countries in which they are most pronounced?

Theory presented in this chapter suggests that if wages are above market-clearing levels, unemployment will result (the number of people seeking work

will exceed the number of available jobs). Furthermore, if wages are held above market-clearing levels and the labor demand curve *shifts to the left*, unemployment will rise to even higher levels (you should be able to show this by drawing a graph with an unchanging supply curve, a fixed wage rate, and a leftward-shifting demand curve). Moreover, above-market wages deter the growth of *new* jobs, so wages “stuck” above market-clearing levels also can cause those who suffer a spell of unemployment to remain in that status for a long time. Thus, measures of the incidence and duration of unemployment—which, fortunately, are comparably defined and estimated in several advanced economies—can sometimes be used to infer the relative strength of nonmarket forces across countries. Consider, for example, what happened to unemployment rates in Europe and North America in the 1980s and 1990s.

One phenomenon characterizing the 1980s was an acceleration of technological change, associated primarily with computerization, in the advanced economies of the world. These changes led to a fall in the demand for less-skilled, less-educated, lower-paid workers. In Canada and the United States the decline in demand for low-skilled workers led to a fall in their real wages throughout the 1980s; despite that, the unemployment rate for less-educated workers rose over that decade—from 7.2 percent to 8.5 percent in the United States and from 6.3 percent to 9.3 percent in Canada. In the two European countries for which we have data on wages and unemployment by skill level, however, the real wages of low-paid workers *rose* over the decade, with the consequence that increases in unemployment for the less educated were much more pronounced. In France, real wages among the lowest-paid workers rose 1 percent per year, and their unemployment rate increased from 4.6 percent to 10.7 percent over the decade. In Germany, where the pay of low-wage workers rose an average of 5 percent per year, unemployment rates among these workers went from 4.4 percent to 13.5 percent.¹⁰

Evidence that nonmarket forces are probably stronger in most of Europe than in North America can be seen in Table 2.4, which compares unemployment rates across countries. While overall rates are not systematically different, the percentages unemployed for longer than one year are generally greater in Europe. Later, we will identify some of the nonmarket forces that might be responsible.¹¹

¹⁰Earnings data for all four countries are for workers in the lowest decile (lowest 10 percent) of their country's earnings distribution. These data are found in Organisation for Economic Co-operation and Development (OECD), *Employment Outlook: July 1993* (Paris: OECD, 1993), Table 5.3. Data on unemployment rates are from Federal Reserve Bank of Kansas City, *Reducing Unemployment: Current Issues and Policy Options* (Kansas City, Mo.: Federal Reserve Bank of Kansas City, 1994): 25.

¹¹For analyses of the relative performance of labor markets in Europe and the United States, see Francine D. Blau and Lawrence M. Kahn, *At Home and Abroad: U.S. Labor-Market Performance in International Perspective* (New York: Russell Sage Foundation, 2002); Gilles Saint-Paul, “Why Are European Countries Diverging in Their Unemployment Experience?” *Journal of Economic Perspectives* 18 (Fall 2004): 49–68; Richard Freeman, *America Works: The Exceptional U.S. Labor Market* (New York: Russell Sage Foundation, 2007); and Stephen Nickell, “Is the U.S. Labor Market Really that Exceptional? A Review of Richard Freeman's *America Works: The Exceptional U.S. Labor Market*,” *Journal of Economic Literature* 46 (June 2008): 384–395.

Table 4.4**Unemployment and Long-Term Unemployment, Selected European and North American Countries, 2007**

	Unemployment Overall Rate	Percent of Unemployed Out of Work > One Year	Unemployment Long-Term Rate
Belgium	7.5%	50.0%	3.8%
Canada	6.0	7.5	0.5
Denmark	3.8	18.2	0.7
France	8.3	40.4	3.4
Germany	8.4	56.6	4.8
Ireland	4.6	30.3	1.4
Netherlands	3.2	41.7	1.3
Norway	2.5	8.5	0.2
United Kingdom	5.3	24.5	1.3
United States	4.6	10.0	0.5

Source: OECD, *Employment Outlook* (Paris: OECD, 2009), Tables A and G.

Review Questions

- As discussed on page 45, in the early 1970s, Egypt experienced a dramatic outflow of construction workers seeking higher wages in Saudi Arabia at the same time that the demand for their services rose within Egypt. Graphically represent these two shifts of supply and demand, and then use the graph to predict the direction of change in wages and employment within Egypt's construction sector during that period.
- Analyze the impact of the following changes on wages and employment in a given occupation:
 - A decrease in the danger of the occupation.
 - An increase in product demand.
 - Increased wages in alternative occupations.
- What would happen to the wages and employment levels of engineers if government expenditures on research and development programs were to fall? Show the effect graphically.
- Suppose a particular labor market were in market-clearing equilibrium. What could happen to cause the equilibrium wage to fall? Suppose price levels were rising each year, but money wages were "sticky downward" and never fell; how would real wages in this market adjust?
- Assume that you have been hired by a company to do a salary survey of its arc welders, who the company suspects are overpaid. Given the company's expressed desire to maximize profits, what definition of *overpaid* would you apply in this situation, and how would you identify whether arc welders are, in fact, overpaid?
- Ecuador is the world's leading exporter of bananas, which are grown and harvested by a large labor force that includes many children. Assume Ecuador now outlaws the use of child labor on banana plantations. Using economic theory in its positive mode, analyze what would happen to employment and wages in the banana

- farming industry in Ecuador. Use supply and demand curves in your analysis.
7. Unions can raise wages paid to their members in two ways. (i) Unions can negotiate a wage rate that lies above the market-clearing wage. While management cannot pay below that rate, management does have the right to decide how many workers to hire. (ii) Construction unions often have agreements that require management to hire only union members, but they also have the power to control entry into the union. Hence, they can raise wages by restricting labor supply.
 - a. Graphically depict method (i) above using a labor supply and a labor demand curve. Show the market-clearing wage as W_e , the market-clearing employment level as L_e , the (higher) negotiated wage as W_u , the level of employment associated with W_u as L_u , and the number of workers wanting to work at W_u as L_s .
 - b. Graphically depict method (ii) above using a labor supply and a labor demand curve. Show the market-clearing wage as W_e , the market-clearing employment level as L_e , the number of members the union decides to have as L_u (which is less than L_e), and the wage associated with L_u as W_u .
 8. American students have organized opposition to the sale by their campus stores of university apparel made for American retailers by workers in foreign countries who work in sweatshop conditions (long hours at low pay in bad working conditions). Assume this movement takes the form of boycotting items made under sweatshop conditions.
 - a. Analyze the immediate labor market outcomes for sweatshop workers in these countries using supply and demand curves to illustrate the mechanisms driving the outcomes.
 - b. Assuming that actions by American students are the only force driving the improvement of wages and working conditions in foreign countries, what must these actions include to ensure that the workers they are seeking to help are unambiguously better off?
 9. Suppose the Occupational Safety and Health Administration were to mandate that all punch presses be fitted with a very expensive device to prevent injuries to workers. This device does not improve the efficiency with which punch presses operate. What does this requirement do to the demand curve for labor? Explain.
 10. Suppose we observe that employment levels in a certain region suddenly decline as a result of (i) a fall in the region's demand for labor and (ii) wages that are fixed in the short run. If the *new* labor demand curve remains unchanged for a long period and the region's labor supply curve does not shift, is it likely that employment in the region will recover? Explain.
 11. In the economic recovery of 2003–2004, job growth in Canada was much faster than job growth in the United States. Please answer the following questions: (a) Generally speaking, how does economic growth affect the demand curve for labor? (b) Assume that growth does not affect the labor supply curve in either country, and suppose that the faster job growth in Canada was accompanied by slower (but positive) wage growth there than in the United States. What would this fact tell us about the reasons for the relatively faster job growth in Canada?
 12. Assume that the war in Iraq increased the desired size of the U.S. military, and assume that potential recruits are reduced by the prospect of facing dangerous, unpleasant wartime conditions. First, analyze how the war affects the supply

curve and the demand curve for military personnel. Second, use your analysis to predict how the war will affect the wages

and the employment level of military personnel.

Problems

- Suppose that the adult population is 210 million, and there are 130 million who are employed and 5 million who are unemployed. Calculate the unemployment rate and the labor force participation rate.
- Suppose that the supply curve for schoolteachers is $L_S = 20,000 + 350W$, and the demand curve for schoolteachers is $L_D = 100,000 - 150W$, where L = the number of teachers and W = the daily wage.
 - Plot the supply and demand curves.
 - What are the equilibrium wage and employment levels in this market?
 - Now suppose that at any given wage, 20,000 more workers are willing to work as schoolteachers. Plot the new supply curve, and find the new wage and employment level. Why doesn't employment grow by 20,000?
- Have the real average hourly earnings for production and nonsupervisory workers in the United States risen during the past 12 months? Go to the Bureau of Labor Statistics Web site (<http://stats.bls.gov>) to find the numbers needed to answer the question.
- Suppose the adult population of a city is 9,823,000 and there are 3,340,000 people who are not in the labor force and 6,094,000 who are employed.
 - Calculate the number of adults who are in the labor force and the number of adults who are unemployed.
 - Calculate the labor force participation rate and the unemployment rate.
- From Table 2.2, the CPI (with a base of 100 in 1982–1984) rose from 130.7 in 1990 to 201.6 in 2006. The federal minimum wage (nominal hourly wage) in 1990 was \$3.80, and it was \$5.15 in 2006. Calculate the minimum wage in real (1982–1984) dollars. Did the federal minimum wage increase or decrease in real dollars from 1990 to 2006?
- The following table gives the demand and supply for cashiers in retail stores.

Wage Rate (\$)	Number of Cashiers Demanded	Number of Cashiers Supplied
3.00	200	70
4.00	180	100
5.00	170	120
6.00	150	150
7.00	130	160
8.00	110	175
9.00	80	190

 - Plot the supply and demand curves.
 - What are the equilibrium wage and employment levels in this market?
 - Suppose the number of cashiers demanded increases by 30 at every wage rate. Plot the new demand curve. What are the equilibrium wage and employment level now?
- From the original demand function in Problem 6 (see table), how many cashiers would have jobs if the wage paid were \$8.00 per hour? Discuss the implications of an \$8 wage in the market for cashiers.

Selected Readings

- Blau, Francine D., and Lawrence M. Kahn. *At Home and Abroad: U.S. Labor-Market Performance in International Perspective*. New York: Russell Sage Foundation, 2002.
- Freeman, Richard. *America Works: The Exceptional U.S. Labor Market*. New York: Russell Sage Foundation, 2007.
- Nickell, Stephen. "Is the U.S. Labor Market Really That Exceptional? A Review of Richard Freeman's *America Works: The Exceptional U.S. Labor Market*." *Journal of Economic Literature* 46 (June 2008): 384–395.
- President's Commission on an All-Volunteer Armed Force. *Report of the President's Commission on an All-Volunteer Armed Force*. Chapter 3, "Conscription Is a Tax," 23–33. Washington, D.C.: U.S. Government Printing Office, February 1970.
- Rottenberg, Simon. "On Choice in Labor Markets." *Industrial and Labor Relations Review* 9 (January 1956): 183–199. [Robert J. Lampman. "On Choice in Labor Markets: Comment." *Industrial and Labor Relations Review* 9 (July 1956): 636–641.]
- Saint-Paul, Gilles. "Why Are European Countries Diverging in Their Unemployment Experience?" *Journal of Economic Perspectives* 18 (Fall 2004): 49–68.

CHAPTER 5

The Demand for Labor

The demand for labor is a derived demand, in that workers are hired for the contribution they can make toward producing some good or service for sale. However, the wages workers receive, the employee benefits they qualify for, and even their working conditions are all influenced, to one degree or another, by the government. There are minimum wage laws, pension regulations, restrictions on firing workers, safety requirements, immigration controls, and government-provided pension and unemployment benefits that are financed through employer payroll taxes. All these requirements and regulations have one thing in common: they increase employers' costs of hiring workers.

We explained in chapter 2 that both the scale and the substitution effects accompanying a wage change suggest that the demand curve for labor is a *downward-sloping function of the wage rate*. If this rather simple proposition is true, then policies that mandate increases in the costs of employing workers will have the undesirable side effect of reducing their employment opportunities. If the reduction is large enough, lost job opportunities could actually undo any help provided to workers by the regulations. Understanding the characteristics of labor demand curves, then, is absolutely crucial to anyone interested in public policy. To a great extent, how one feels about many labor market regulatory programs is a function of one's beliefs about labor demand curves!

This chapter will identify *assumptions* underlying the proposition that labor demand is a downward-sloping function of the wage rate. Chapter 4 will take the downward-sloping nature of labor demand curves as given, addressing instead why, in the face of a given wage increase, declines in demand might be large in some cases and barely perceptible in others.

Profit Maximization

The fundamental assumption of labor demand theory is that firms—the employers of labor—seek to maximize profits. In doing so, firms are assumed to continually ask, “Can we make changes that will improve profits?” Two things should be noted about this constant search for enhanced profits. First, a firm can make changes only in variables that are within its control. Because the price a firm can charge for its product and the prices it must pay for its inputs are largely determined by others (the “market”), profit-maximizing decisions by a firm mainly involve the question of *whether, and how, to increase or decrease output*.

Second, because the firm is assumed to constantly search for profit-improving possibilities, our theory must address the *small* (“marginal”) changes that must be made almost daily. Really major decisions of whether to open a new plant or introduce a new product line, for example, are relatively rare; once having made them, the employer must approach profit maximization incrementally through the trial-and-error process of small changes. We therefore need to understand the basis for these incremental decisions, paying particular attention to when an employer *stops* making changes in output levels or in its mix of inputs.

(With respect to the employment of inputs, it is important to recognize that analyzing marginal changes implies considering a small change in one input *while holding employment of other inputs constant*. Thus, when analyzing the effects of adjusting the labor input by one unit, for example, we will do so on the assumption that capital is held constant. Likewise, marginal changes in capital will be considered assuming the labor input is held constant.)

In incrementally deciding on its optimal level of *output*, the profit-maximizing firm will want to expand output by one unit if the added revenue from selling that unit is greater than the added cost of producing it. As long as the marginal revenue from an added unit of output exceeds its marginal cost, the firm will continue to expand output. Likewise, the firm will want to contract output whenever the marginal cost of production exceeds marginal revenue. Profits are maximized (and the firm stops making changes) when output is such that marginal revenue equals marginal cost.

A firm can expand or contract output, of course, only by altering its use of *inputs*. In the most general sense, we will assume that a firm produces its output by combining two types of inputs, or *factors of production: labor and capital*. Thus, the rules stated earlier for deciding whether to marginally increase or reduce output have important corollaries with respect to the employment of labor and capital:

- a. If the income generated by employing one more unit of an input exceeds the additional expense, then add a unit of that input.

- b. If the income generated by one more unit of input is less than the additional expense, reduce employment of that input.
- c. If the income generated by one more unit of input is equal to the additional expense, no further changes in that input are desirable.

Decision rules (a) through (c) state the profit-maximizing criterion in terms of *inputs* rather than output; as we will see, these rules are useful guides to deciding *how*—as well as *whether*—to marginally increase or decrease output. Let us define and examine the components of these decision rules more closely.

Marginal Income from an Additional Unit of Input

Employing one more unit of either labor or capital generates additional income for the firm because of the added output that is produced and sold. Similarly, reducing the employment of labor or capital reduces a firm's income flow because the output available for sale is reduced. Thus, the marginal income associated with a unit of input is found by multiplying two quantities: the change in physical output produced (called the input's *marginal product*) and the MR generated per unit of physical output. We will therefore call the marginal income produced by a unit of input the input's *marginal revenue product*. For example, if the presence of a tennis star increases attendance at a tournament by 20,000 spectators, and the organizers net \$25 from each additional fan, the marginal income produced by this star is equal to her marginal product (20,000 fans) times the marginal revenue of \$25 per fan. Thus, her marginal revenue product equals \$500,000. (For an actual calculation of marginal revenue product in college football, see Example 3.1.)

Marginal Product Formally, we will define the *marginal product of labor*, or MP_L , as the change in physical output (ΔQ) produced by a change in the units of labor (ΔL), holding capital constant:¹

$$MP_L = \Delta Q / \Delta L \quad (\text{holding capital constant}) \quad (3.1)$$

Likewise, the marginal product of capital (MP_K) will be defined as the change in output associated with a one-unit change in the stock of capital (ΔK), holding labor constant:

$$MP_K = \Delta Q / \Delta K \quad (\text{holding capital constant}) \quad (3.2)$$

Marginal Revenue The definitions in equations (3.1) and (3.2) reflect the fact that a firm can expand or contract its output only by increasing or decreasing its use of either labor or capital. The marginal revenue that is generated by an extra unit of output depends on the characteristics of the product market in which that output is

¹The symbol Δ (the uppercase Greek letter delta) is used to signify "a change in."

EXAMPLE 5.1**The Marginal Revenue Product of College Football Stars**

Calculating a worker's marginal revenue product is often very complicated due to lack of data and the difficulty of making sure that everything else is being held constant and only *additions* to revenue are counted. Perhaps for this reason, economists have been attracted to the sports industry, which generates so many statistics on player productivity and team revenues.

Football is a big-time concern on many campuses, and some star athletes generate huge revenues for their colleges, even though they are not paid—except by receiving a free education. Robert Brown collected revenue statistics for 47 Division I-A college football programs for the 1988–1989 season—including revenues retained by the school from ticket sales, donations to the athletic department, and television and radio payments. (Unfortunately, this leaves out some

other potentially important revenue sources, such as parking and concessions at games and donations to the general fund.)

Next, he examined variation in revenues due to market size, strength of opponents, national ranking, and the number of players on the team who were so good that they were drafted into professional football (the National Football League [NFL]). Brown found that each additional player drafted into the NFL was worth about \$540,000 (\$934,000 in 2009 dollars) in extra revenue to his team. Over a four-year college career, a premium player could therefore generate over \$3 million in revenues for his university!

*Data from: Robert W. Brown, "An Estimate of the Rent Generated by a Premium College Football Player," *Economic Inquiry* 31 (October 1993), 671–684.*

sold. If the firm operates in a purely competitive product market, and therefore has many competitors and no control over product price, the marginal revenue per unit of output sold is equal to product price (P). If the firm has a differentiated product, and thus has some degree of monopoly power in its product market, extra units of output can be sold only if product price is reduced (because the firm faces the *market* demand curve for its particular product); students will recall from introductory economics that in this case, marginal revenue is less than price ($MR < P$).²

Marginal Revenue Product Combining the definitions presented in this section, the firm's marginal revenue product of labor, or MRP_L , can be represented as

$$MRP_L = MP_L \cdot MR \quad (\text{in the general case}) \quad (3.3a)$$

or as

$$MRP_L = MP_L \cdot P \quad (\text{if the product market is competitive}) \quad (3.3b)$$

²A competitive firm can sell added units of output at the market price because it is so small relative to the entire market that its output does not affect price. A monopolist, however, *is* the supply side of the product market, so to sell extra output, it must lower price. Because it must lower price on *all* units of output, and not just on the extra units to be sold, the MR associated with an additional unit is below price.

Likewise, the firm's marginal revenue product of capital (MRP_K) can be represented as $MP_K \cdot MR$ in the general case or as $MP_K \cdot P$ if the product market is competitive.

Marginal Expense of an Added Input

Changing the levels of labor or capital employed, of course, will add to or subtract from the firm's total costs. The marginal expense of labor (ME_L) that is incurred by hiring more labor is affected by the nature of competition in the labor market. If the firm operates in a competitive labor market and has no control over the wages that must be paid (it is a "wage taker"), then ME_L is simply equal to the market wage. Put differently, firms in competitive labor markets have labor supply curves that are horizontal at the going wage (refer back to Figure 2.11); if they hire an additional hour of labor, their costs increase by an amount equal to the wage rate, W .

In this chapter, we will maintain the assumption that the labor market is competitive and that the labor supply curve to firms is therefore *horizontal* at the going wage. In chapter 5, we will relax this assumption and analyze how upward-sloping labor supply curves to individual employers alter the marginal expense of labor.

In the analysis that follows, the marginal expense of adding a unit of capital will be represented as C , which can be thought of as the expense of renting a unit of capital for one time period. The specific calculation of C need not concern us here, but it clearly depends on the purchase price of the capital asset, its expected useful life, the rate of interest on borrowed funds, and even special tax provisions regarding capital.

The Short-Run Demand for Labor When Both Product and Labor Markets Are Competitive

The simplest way to understand how the profit-maximizing behavior of firms generates a labor demand curve is to analyze the firm's behavior over a period of time so short that the firm cannot vary its stock of capital. This period is what we will call the *short run*, and, of course, the time period involved will vary from firm to firm (an accounting service might be able to order and install a new computing system for the preparation of tax returns within three months, whereas it may take an oil refinery five years to install a new production process). What is simplifying about the short run is that, with capital fixed, a firm's choice of output level and its choice of employment level are two aspects of the very same decision. Put differently, in the short run, the firm needs only to decide *whether* to alter its output level; *how* to increase or decrease output is not an issue, because only the employment of labor can be adjusted.

Table 5.1

The Marginal Product of Labor in a Hypothetical Car Dealership (Capital Held Constant)

Number of Salespersons	Total Cars Sold	Marginal Product of Labor
0	0	10
1	10	11
2	21	5
3	26	3
4	29	

A Critical Assumption: Declining MP_L

We defined the marginal product of labor MP_L as the change in the (physical) output of a firm when it changes its employment of labor by one unit, holding capital constant. Since the firm can vary its employment of labor, we must consider how increasing or reducing labor will affect labor's marginal product. Consider Table 3.1, which illustrates a hypothetical car dealership with sales personnel who are all equally hardworking and persuasive. With no sales staff, the dealership is assumed to sell zero cars, but with one salesperson, it will sell 10 cars per month. Thus, the marginal product of the first salesperson hired is 10. If a second person is hired, total output is assumed to rise from 10 to 21, implying that the marginal product of a second salesperson is 11. If a third equally persuasive salesperson is hired, sales rise from 21 to 26 ($MP_L = 5$), and if a fourth is hired, sales rise from 26 to 29 ($MP_L = 3$).

Table 3.1 assumes that adding an extra salesperson increases output (cars sold) in each case. As long as output *increases* as labor is added, labor's marginal product is *positive*. In our example, however, MP_L increased at first (from 10 to 11) but then fell (to 5 and eventually to 3). Why?

The initial rise in marginal product occurs *not* because the second salesperson is better than the first; we ruled out this possibility by our assumption that the salespeople were equally capable. Rather, the rise could be the result of cooperation between the two in generating promotional ideas or helping each other out in some way. Eventually, however, as more salespeople are hired, MP_L must fall. A fixed building (remember that capital is held constant) can contain only so many cars and customers; thus, each additional increment of labor must eventually produce progressively smaller increments of output. This law of *diminishing marginal returns* is an empirical proposition that derives from the fact that as employment expands, each additional worker has a progressively smaller share of the capital stock to work with. For expository convenience, we shall assume that MP_L is always decreasing.³

³We lose nothing by this assumption because we show later in this section that a firm will never be operated at a point where its MP_L is increasing.

From Profit Maximization to Labor Demand

From the profit-maximizing decision rules discussed earlier, it is clear that the firm should keep increasing its employment of labor as long as labor's marginal revenue product exceeds its marginal expense. Conversely, it should keep reducing its employment of labor as long as the expense saved is greater than the income lost. *Profits are maximized, then, only when employment is such that any further one-unit change in labor would have a marginal revenue product equal to marginal expense:*

$$MRP_L = ME_L \quad (3.4)$$

Under our current assumptions of competitive product and labor markets, we can symbolically represent the profit-maximizing level of labor input as that level at which

$$MP_L \cdot P = W \quad (3.5)$$

Clearly, equation (3.5) is stated in terms of some *monetary* unit (dollars, for example).

Alternatively, however, we can divide both sides of equation (3.5) by product price, P , and state the profit-maximizing condition for hiring labor in terms of *physical quantities*:

$$MP_L = W/P \quad (3.6)$$

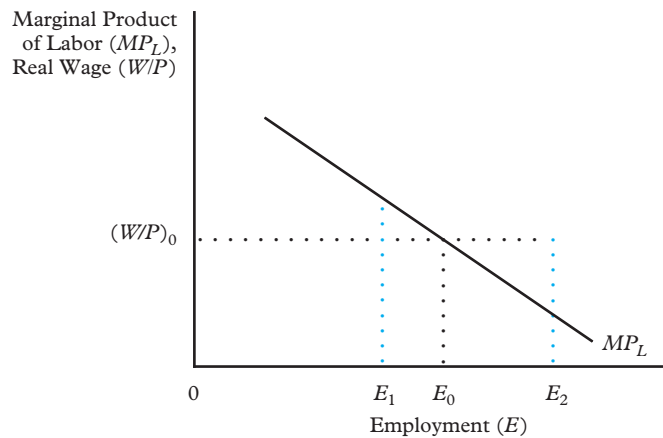
We defined MP_L as the change in physical output associated with a one-unit change in labor, so it is obvious that the left-hand side of equation (3.6) is in physical quantities. To understand that the right-hand side is also in physical quantities, note that the numerator (W) is the dollars per unit of labor, and the denominator (P) is the dollars per unit of output. Thus, the ratio W/P has the dimension of physical units. For example, if a woman is paid \$10 per hour and the output she produces sells for \$2 per unit, from the firm's viewpoint, she is paid five units of output per hour ($10 \div 2$). From the perspective of the firm, these five units represent her "real wage."

Labor Demand in Terms of Real Wages The demand for labor can be analyzed in terms of either *real* or *money* wages. Which version of demand analysis is used is a matter of convenience only. In this and the following section, we give examples of both.

Figure 3.1 shows a marginal product of labor (MP_L) schedule for a representative firm. In this figure, the MP_L is tabulated on the vertical axis and the number of units of labor employed on the horizontal axis. The negative slope of the schedule indicates that each additional unit of labor employed produces a progressively smaller (but still positive) increment in output. Because the real wage and MP_L are both measured in the same dimension (units of output), we can also plot the real wage on the vertical axis of Figure 3.1.

Figure 5.1

Demand for Labor in the Short Run
(Real Wage)



Given any real wage (by the market), the firm should thus employ labor to the point at which MP_L just equals the real wage (equation 3.6). In other words, *the firm's demand for labor in the short run is equivalent to the downward-sloping segment of its MP_L schedule.*⁴

To see that this is true, pick any real wage—for example, the real wage denoted by $(W/P)_0$ in Figure 3.1. We have asserted that the firm's demand for labor is equal to its MP_L schedule and, consequently, that the firm would employ E_0 employees. Now, suppose that a firm initially employed E_2 workers as indicated in Figure 3.1, where E_2 is *any* employment level greater than E_0 . At the employment level E_2 , the MP_L is less than the real wage rate; the marginal real cost of the last unit of labor hired is therefore greater than its marginal product. As a result, profit could be increased by reducing the level of employment. Similarly, suppose instead that a firm initially employed E_1 employees, where E_1 is *any* employment level less than E_0 . Given the specified real wage $(W/P)_0$, the MP_L is greater than the real wage rate at E_1 —and, consequently, the marginal additions to output of an extra unit of labor exceed its marginal real cost. As a result, a firm could increase its profit level by expanding its level of employment.

Hence, to maximize profits, given any real wage rate, a firm should stop employing labor at the point at which any additional labor would cost more than it would produce. This profit-maximization rule implies two things. First, the firm should employ labor up to the point at which its real wage equals MP_L —but not beyond that point.

⁴We should add here, “provided that the firm's revenue exceeds its labor costs.” Above some real wage level, this may fail to occur, and the firm will go out of business (employment will drop to zero).

Second, its profit-maximizing level of employment lies in the range where its MP_L is *declining*. if $W/P = MP_L$, but MP_L is *increasing*, then adding another unit of labor will create a situation in which marginal product *exceeds* W/P . As long as adding labor causes MP_L to exceed W/P , the profit-maximizing firm will continue to hire labor. It will stop hiring only when an extra unit of labor would reduce MP_L below W/P , which will happen only when MP_L is declining. Thus, the only employment levels that could possibly be consistent with profit maximization are those in the range where MP_L is decreasing.

Labor Demand in Terms of Money Wages In some circumstances, labor demand curves are more readily conceptualized as downward-sloping functions of *money* wages. To make the analysis as concrete as possible, in this section, we analyze the demand for department store detectives.

At a business conference one day, a department store executive boasted that his store had reduced theft to 1 percent of total sales. A colleague shook her head and said, "I think that's too low. I figure it should be about 2 percent of sales." How can more shoplifting be better than less? The answer is based on the fact that reducing theft is costly in itself. A profit-maximizing firm will not want to take steps to reduce shoplifting if the added costs it must bear in so doing exceed the value of the savings such steps will generate.

Table 3.2 shows a hypothetical marginal revenue product of labor MRP_L schedule for department store detectives. Hiring one detective would, in this example, save \$50 worth of thefts per hour. Two detectives could save \$90 worth of thefts each hour, or \$40 more than hiring just one. The MRP_L of hiring a second detective is thus \$40. A third detective would add \$20 more to thefts prevented each hour.

The MRP_L does *not* decline from \$40 to \$20 because the added detectives are incompetent; in fact, we shall assume that all are equally alert and well trained. MRP_L declines, in part, because surveillance equipment (capital) is fixed; with each added detective, there is less equipment per person. However, the MRP_L also declines because it becomes progressively harder to generate savings. With just a few detectives, the only thieves caught will be the more-obvious, less-experienced

Table 5.2**Hypothetical Schedule of Marginal Revenue Productivity of Labor for Store Detectives**

Number of Detectives on Duty during Each Hour Store Is Open	Total Value of Thefts Prevented per Hour	Marginal Value of Thefts Prevented per Hour (MRP_L)
0	\$ 0	\$—
1	\$ 50	\$50
2	\$ 90	\$40
3	\$110	\$20
4	\$115	\$ 5
5	\$117	\$ 2

shoplifters. As more detectives are hired, it becomes possible to prevent theft by the more-expert shoplifters, but they are harder to detect and fewer in number. Thus, MRP_L falls because theft prevention becomes more difficult once all those who are easy to catch are apprehended.

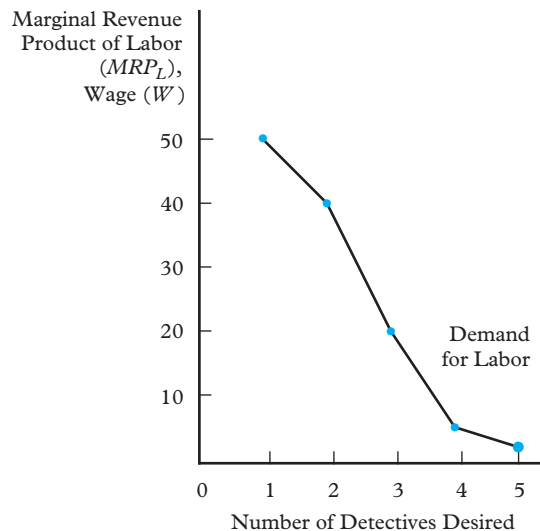
To draw the demand curve for labor, we need to determine how many detectives the store will want to hire at any given wage rate, keeping in mind that employers—through part-time employment—are able to hire fractional workers. For example, at a wage of \$50 per hour, how many detectives will the store want? Using the $MRP_L = W$ criterion (equation 3.5), the answer is “up to one.” At \$40 per hour, the store would want to stop hiring at two, and at \$20 per hour, it would stop at three. The labor demand curve that summarizes the store’s profit-maximizing employment of detectives is shown in Figure 3.2.

Figure 3.2 illustrates a fundamental point: the labor demand curve in the short run slopes downward because it *is* the MRP_L curve—and the MRP_L curve slopes downward because of labor’s diminishing marginal product. The demand curve and the MRP_L curve coincide; this could be demonstrated by graphing the MRP_L schedule in Table 3.2, which would yield exactly the same curve as in Figure 3.2. When one detective is hired, MRP_L is \$50; when two are hired, MRP_L is \$40; and so forth. Since MRP_L always equals W for a profit maximizer who takes wages as given, the MRP_L curve and labor demand curve (expressed as a function of the money wage) must be the same.

An implication of our example is that there is some level of shoplifting the store finds more profitable to tolerate than to eliminate. This level will be higher at high wages for store detectives than at lower wages. To say the theft rate is “too

Figure 5.2

Demand for Labor in the Short Run (Money Wage)



low” thus implies that the marginal costs of crime reduction exceed the marginal savings generated, and the firm is therefore failing to maximize profits.

Finally, we must emphasize that the marginal product of an individual is *not* a function solely of his or her personal characteristics. As stressed earlier, the marginal product of a worker depends upon the number of similar employees the firm has already hired. An individual’s marginal product also depends upon the size of the firm’s capital stock; increases in the firm’s capital stock shift the entire MP_L schedule up. It is therefore incorrect to speak of an individual’s productivity as an immutable factor that is associated only with his or her characteristics, independent of the characteristics of the other inputs he or she has to work with.

Market Demand Curves The demand curve (or schedule) for an individual firm indicates how much labor that firm will want to employ at each wage level. A *market demand curve* (or schedule) is just the *summation* of the labor demanded by all firms in a particular labor market at each level of the *real* wage.⁵ If there are three firms in a certain labor market, and if at a *given* real wage firm A wants 12 workers, firm B wants 6, and firm C wants 20, then the market demand at that real wage is 38 employees. More important, because market demand curves are so closely derived from firm demand curves, they too will *slope downward* as a function of the real wage. When the real wage falls, the number of workers that existing firms want to employ increases. In addition, the lower real wage may make it profitable for new firms to enter the market. Conversely, when the real wage increases, the number of workers that existing firms want to employ decreases, and some firms may be forced to cease operations completely.

Objections to the Marginal Productivity Theory of Demand Two kinds of objections are sometimes raised to the theory of labor demand introduced in this section. The first is that almost no employer can ever be heard uttering the words “marginal revenue product of labor” and that the theory assumes a degree of sophistication that most employers do not have. Employers, it is also argued, are unable in many situations to accurately measure the output of individual workers.

These first objections can be answered as follows: Whether employers can verbalize the profit-maximizing conditions or whether they can explicitly measure the MRP_L , they must at least *intuit* them to survive in a competitive environment. Competition will “weed out” employers who are not good at generating profits, just as competition will weed out pool players who do not understand the intricacies of how speed, angles, and spin affect the motion of bodies through space. Yet, one could canvass the pool halls of America and probably find few who could verbalize Newton’s laws of motion! The point is

⁵If firms’ demand curves are drawn as a function of the money wage, they represent the downward-sloping portion of the firms’ MRP_L curves. In a competitive industry, the price of the product is given to the firm by the market; thus, at the firm level, the MRP_L has imbedded in it a given product price. When aggregating labor demand to the *market* level, product price can no longer be taken as given, and the aggregation is no longer a simple summation.

that employers can *know* concepts without being able to verbalize them. Those that are not good at maximizing profits will not last very long in competitive markets.

The second objection is that in many cases, it seems that adding labor while holding capital constant would not add to output at all. For example, one secretary and one computer can produce output, but it might seem that adding a second secretary while holding the number of computers constant could produce nothing extra, since that secretary would have no machine on which to work.

The answer to this second objection is that the two secretaries could take turns using the computer so that neither became fatigued to the extent that mistakes increased and typing speeds slowed down. The second secretary could also answer the telephone and expedite work in other ways. Thus, even with technologies that seem to require one machine per person, labor will generally have a marginal product greater than zero if capital is held constant.

The Demand for Labor in Competitive Markets When Other Inputs Can Be Varied

An implication of our theory of labor demand is that, because labor can be varied in the short run—that is, at any time—the profit-maximizing firm will always operate so that labor’s marginal revenue product equals the wage rate (which is labor’s marginal expense in a competitive labor market). What we must now consider is how the firm’s ability to adjust *other* inputs affects the demand for labor. We first analyze the implications of being able to adjust capital in the long run, and we then turn our attention to the case of more than two inputs.

Labor Demand in the Long Run

To maximize profits in the long run, the firm must adjust both labor and capital so that the marginal revenue product of each equals its marginal expense. Using the definitions discussed earlier in this chapter, profit maximization requires that the following two equalities be satisfied:

$$MP_L \cdot P = W \quad (\text{a restatement of equation 3.5}) \quad (3.7a)$$

$$MP_K \cdot P = C \quad (\text{the profit-maximizing condition for capital}) \quad (3.7b)$$

Equations (3.7a) and (3.7b) can be rearranged to isolate P , so these two profit-maximizing conditions can also be expressed as

$$P = W/MP_L \quad (\text{a rearrangement of equation 3.7a}) \quad (3.8a)$$

$$P = C/MP_K \quad (\text{a rearrangement of equation 3.7b}) \quad (3.8b)$$

Furthermore, because the right-hand sides of equations (3.8a) and (3.8b) equal the same quantity, P , profit maximization therefore requires that

$$W/MP_L = C/MP_K \quad (3.8c)$$

The economic meaning of equation (3.8c) is key to understanding how the ability to adjust capital affects the firm's demand for labor. Consider the left-hand side of equation (3.8c): the numerator is the cost of a unit of labor, while the denominator is the extra output produced by an added unit of labor. Therefore, the ratio W/MP_L turns out to be the added cost of producing an added unit of output when using labor to generate the increase in output.⁶ Analogously, the right-hand side is the marginal cost of producing an extra unit of output using capital. What equation (3.8c) suggests is that to maximize profits, *the firm must adjust its labor and capital inputs so that the marginal cost of producing an added unit of output using labor is equal to the marginal cost of producing an added unit of output using capital.* Why is this condition a requirement for maximizing profits?

To maximize profits, a firm must be producing its chosen level of output in the least-cost manner. Logic suggests that as long as the firm can expand output more cheaply using one input than the other, it cannot be producing in the least-cost way. For example, if the marginal cost of expanding output by one unit using labor were \$10, and the marginal cost using capital were \$12, the firm could keep output constant and lower its costs of production! How? It could reduce its capital by enough to cut output by one unit (saving \$12) and then add enough labor to restore the one-unit cut (costing \$10). Output would be the same, but costs would have fallen by \$2. Thus, for the firm to be maximizing profits, it must be operating at the point such that further marginal changes in both labor and capital would neither lower costs nor add to profits.

With equations (3.8a) to (3.8c) in mind, what would happen to the demand for labor in the long run if the wage rate (W) facing a profit-maximizing firm were to rise? First, as we discussed in the section on the "The Short-Run Demand for Labor When Both Product and Labor Markets Are Competitive," the rise in W disturbs the equality in equation (3.8a), and the firm will want to cut back on its use of labor even before it can adjust capital. Because the MP_L is assumed to rise as employment is reduced, any cuts in labor will raise MP_L .

Second, because each unit of capital now has less labor working with it, the MP_K falls, disturbing the equality in equation (3.8b). By itself, this latter inequality will cause the firm to want to reduce its stock of capital.

Third, the rise in W will initially end the equality in equation (3.8c), meaning that the marginal cost of production using labor now exceeds the marginal cost using capital. If the above cuts in labor are made in the short run, the associated increase in MP_L and decrease in MP_K will work toward restoring equality in equation (3.8c);

⁶Because $MP_L = \Delta Q/\Delta L$, the expression W/MP_L can be rewritten as $W \cdot \Delta L/\Delta Q$. Since $W\Delta L$ represents the added cost from employing one more unit of labor, the expression $W\Delta L/\Delta Q$ equals the cost of an added unit of output when that unit is produced by adding labor.

EXAMPLE 5.2**Coal Mining Wages and Capital Substitution**

That wage increases have both a *scale effect* and a *substitution effect*, both of which tend to reduce employment, is widely known—even by many of those pushing for higher wages. John L. Lewis was president of the United Mine Workers from the 1920s through the 1940s, when wages for miners were increased considerably with full knowledge that this would induce the substitution of capital for labor. According to Lewis:

Primarily the United Mine Workers of America insists upon the maintenance of the wage standards guaranteed by the existing contractual relations in the industry, in the interests of its own membership. . . . But in insisting on the maintenance of an American wage standard in the coal fields the United Mine Workers is also doing its part, probably more than its part,

to force a reorganization of the basic industry of the country upon scientific and efficient lines. The maintenance of these rates will accelerate the operation of natural economic laws, which will in time eliminate uneconomic mines, obsolete equipment, and incompetent management.

The policy of the United Mine Workers of America will inevitably bring about the utmost employment of machinery of which coal mining is physically capable. . . . Fair wages and American standards of living are inextricably bound up with the progressive substitution of mechanical for human power. It is no accident that fair wages and machinery will walk hand-in-hand.

Source: John L. Lewis, *The Miners' Fight for American Standards* (Indianapolis: Bell, 1925): 40, 41, 108.

however, if it remains more costly to produce an extra unit of output using labor than using capital, the firm will want to substitute capital for labor in the long run. Substituting capital for labor means that the firm will produce its profit-maximizing level of output (which is clearly reduced by the rise in W) in a more capital-intensive way. The act of substituting capital for labor also will serve to increase MP_L and reduce MP_K , thereby reinforcing the return to equality in equation (3.8c).

In the end, the increase in W will cause the firm to reduce its desired employment level for two reasons. The firm's profit-maximizing level of output will fall, and the associated reduction in required inputs (both capital and labor) is an example of the *scale effect*. The rise in W also causes the firm to substitute capital for labor so that it can again produce in the least-cost manner; changing the mix of capital and labor in the production process is an example of the *substitution effect*. The scale and substitution effects of a wage increase will have an ambiguous effect on the firm's desired stock of *capital*, but both effects serve to reduce the demand for *labor*. Thus, as illustrated in Example 3.2, the long-run ability to adjust capital lends further theoretical support to the proposition that the labor demand curve is a downward-sloping function of the wage rate.

More Than Two Inputs

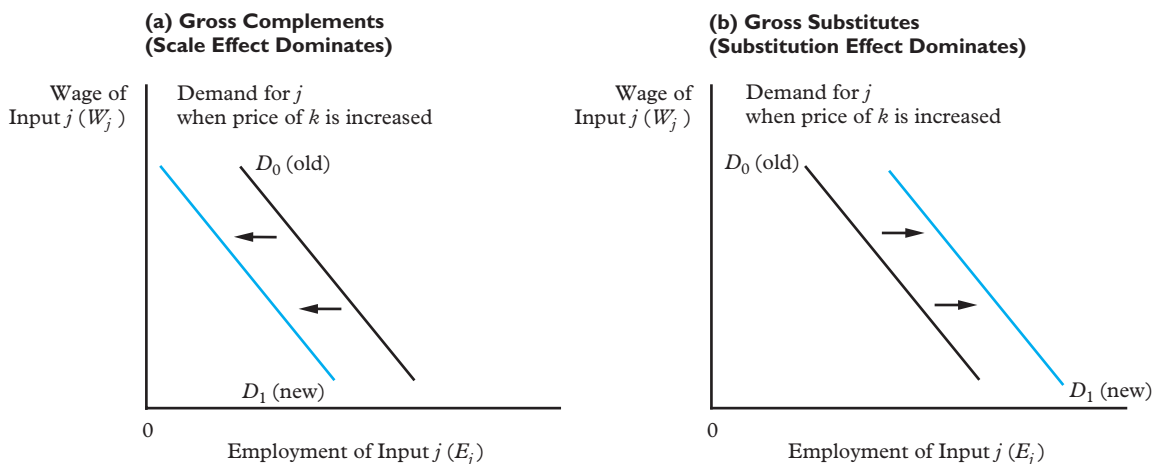
Thus far, we have assumed that there are only two inputs in the production process: capital and labor. In fact, labor can be subdivided into many categories; for example, labor can be categorized by age, educational level, and occupation.

Other inputs that are used in the production process include materials and energy. If a firm is seeking to minimize costs, in the long run, it should employ all inputs up until the point that the marginal cost of producing an added unit of output is the same regardless of which input is increased. This generalization of equation (3.8c) leads to the somewhat obvious result that the demand for *any* category of labor will be a function of its own wage rate *and* (through the scale and substitution effects) the wage or prices of all other categories of labor, capital, and supplies.

If Inputs Are Substitutes in Production The demand curve for each category of labor will be a downward-sloping function of the wage rate paid to workers in that category for the reasons discussed earlier, but how is it affected by wage or price changes for *other* inputs? If two inputs are *substitutes in production* (that is, if the greater use of one in producing output can compensate for reduced use of the other), then increases in the price of the *other* input may shift the entire demand curve for a *given* category of labor either to the right or to the left, depending on the relative strength of the substitution and scale effects. If an increase in the price of one input shifts the demand for *another* input to the left, as in panel (a) of Figure 3.3, the scale effect has dominated the substitution effect, and the two inputs are said to be *gross complements*; if the increase shifts the demand for the other input to the right, as in panel (b) of Figure 3.3, the substitution effect has dominated, and the two inputs are *gross substitutes*.

Figure 5.3

Effect of Increase in the Price of One Input (k) on Demand for Another Input (j), Where Inputs Are Substitutes in Production



If Inputs Are Complements in Production If, instead, the two inputs must be used together—in which case they are called *perfect complements* or *complements in production*—then reduced use of one implies reduced use of the other. In this case, there is no substitution effect, only a scale effect, and the two inputs must be gross complements.

Examples Consider an example of a snow-removal firm in which skilled and unskilled workers are substitutes in production—snow can be removed using either unskilled workers (with shovels) or skilled workers driving snowplows. Let us focus on demand for the skilled workers. Other things equal, an increase in the wage of skilled workers would cause the firm to employ fewer of them; their demand curve would be a downward-sloping function of their wage. If only the wage of *unskilled* workers increased, however, the employer would want fewer unskilled workers than before, and more of the now relatively less-expensive skilled workers, to remove any *given amount of snow*. To the extent that this substitution effect dominated over the scale effect, the demand for skilled workers would shift to the right. In this case, skilled and unskilled workers would be gross substitutes. In contrast, if the reduction in the scale of output caused employment of skilled workers to be reduced, even though skilled workers were being substituted for unskilled workers in the production process, skilled and unskilled workers would be considered gross complements.

In the above firm, snowplows and skilled workers are complements in production. If the price of snowplows went up, the employer would want to cut back on their use, which would result in a reduced demand at each wage for the skilled workers who drove the snowplows. As noted above, inputs that are complements in production are always gross complements.

Labor Demand When the Product Market Is Not Competitive

Our analysis of the demand for labor, in both the short and the long run, has so far taken place under the assumption that the firm operates in competitive product and labor markets. This is equivalent to assuming that the firm is both a price taker and a wage taker; that is, that it takes both P and W as given and makes decisions only about the levels of output and inputs. We will now explore the effects of noncompetitive (monopolistic) *product* markets on the demand for labor (the effects of noncompetitive *labor* markets will be analyzed in chapter 5).

Maximizing Monopoly Profits

As explained earlier in footnote 2 and the surrounding text, product-market monopolies are subject to the *market* demand curve for their output, and they therefore do not take output price as given. They can expand their sales only by

reducing product price, which means that their marginal revenue (MR) from an extra unit of output is less than product price (P). Using the general definition of marginal revenue product in equation (3.3a), and applying the usual profit-maximizing criteria outlined in equation (3.4) to a monopoly that searches for workers in a competitive *labor* market (so that $ME_L = W$), the monopolist would hire workers until its marginal revenue product of labor (MRP_L) equals the wage rate:

$$MRP_L = MR \cdot MP_L = W \quad (3.9)$$

Now we can express the demand for labor in the short run in terms of the real wage by dividing equation (3.9) by the firm's product price, P , to obtain

$$\frac{MR}{P} \cdot MP_L = \frac{W}{P} \quad (3.10)$$

Since marginal revenue is always less than a monopoly's product price, the ratio MR/P in equation (3.10) is less than one. Therefore, the labor demand curve for a firm that has monopoly power in the output market will lie below and to the left of the labor demand curve for an *otherwise identical* firm that takes product price as given. Put another way, just as the level of profit-maximizing output is lower under monopoly than it is under competition, other things equal, so is the level of employment.

The *wage* rates that monopolies pay, however, are not necessarily different from competitive levels even though *employment* levels are. An employer with a product-market monopoly may still be a very small part of the market for a particular kind of employee and thus be a *price taker* in the labor market. For example, a local utility company might have a product-market monopoly, but it would have to compete with all other firms to hire clerks and thus would have to pay the going wage.

Do Monopolies Pay Higher Wages?

Economists have long suspected that product-market monopolies pay wages that are *higher* than what competitive firms would pay.⁷ Monopolies are often regulated by the government to prevent them from exploiting their status and earning monopoly profits, but they are allowed to pass along to consumers their costs of production. Thus, while unable to maximize profits, the managers of a monopoly can enhance their *utility* by paying high wages and passing the costs

⁷For a full statement of this argument, see Armen Alchian and Reuben Kessel, "Competition, Monopoly, and the Pursuit of Money," in *Aspects of Labor Economics*, ed. H. G. Lewis (Princeton, N.J.: Princeton University Press, 1962).

along to consumers in the form of higher prices. The ability to pay high wages makes a manager's life more pleasant by making it possible to hire people who might be more attractive or personable or have other characteristics managers find desirable.

The evidence on monopoly wages, however, is not very clear as yet. Some studies suggest that firms in industries with relatively few sellers *do* pay higher wages than competitive firms for workers with the same education and experience. Other studies of regulated monopolies, however, have obtained mixed results on whether wages tend to be higher for comparable workers in these industries.⁸

Policy Application: The Labor Market Effects of Employer Payroll Taxes and Wage Subsidies

We now apply labor demand theory to the phenomena of employer payroll taxes and wage subsidies. Governments widely finance certain social programs through taxes that require *employers* to remit payments based on their total payroll costs. As we will see, new or increased payroll taxes levied on the employer raise the cost of hiring labor, and they might therefore be expected to reduce the demand for labor. Conversely, it can be argued that if the government were to subsidize the wages paid by employers, the demand for labor would increase; indeed, wage subsidies for particular disadvantaged groups in society are sometimes proposed as a way to increase their employment. In this section, we will analyze the effects of payroll taxes and subsidies.

Who Bears the Burden of a Payroll Tax?

Payroll taxes require employers to pay the government a certain percentage of their employees' earnings, often up to some maximum amount. Unemployment insurance as well as Social Security retirement, disability, and Medicare programs are prominent examples. Does taxing employers to generate revenues for these programs relieve *employees* of a financial burden that would otherwise fall on them?

Suppose that only the employer is required to make payments and that the tax is a fixed amount (X) per labor hour rather than a percentage of payroll.

⁸Ronald Ehrenberg, *The Regulatory Process and Labor Earnings* (New York: Academic Press, 1979); Barry T. Hirsch, "Trucking Regulation, Unionization, and Labor Earnings," *Journal of Human Resources* 23 (Summer 1988): 296–319; S. Nickell, J. Vainiomaki, and S. Wadhvani, "Wages and Product Market Power," *Economica* 61 (November 1994): 457–473; and Marianne Bertrand and Sendhil Mullainathan, "Is There Discretion in Wage Setting? A Test Using Takeover Legislation," *RAND Journal of Economics* 30 (Autumn 1999): 535–554.

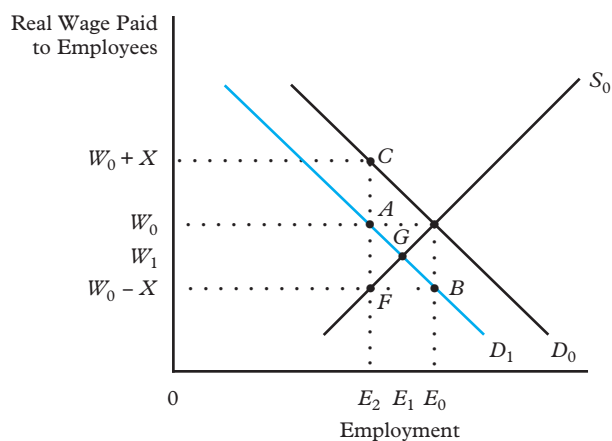
Now, consider the market demand curve D_0 in Figure 3.4, which is drawn in such a way that desired employment is plotted against the wage *employees receive*. Prior to the imposition of the tax, the wage employees receive is the same as the wage employers pay. Thus, if D_0 were the demand curve before the tax was imposed, it would have the conventional interpretation of indicating how much labor firms would be willing to hire at any given wage. However, *after* imposition of the tax, employer wage costs would be X above what employees received.

Shifting the Demand Curve If employees received W_0 , employers would now face costs of $W_0 + X$. They would no longer demand E_0 workers; rather, because their costs were $W_0 + X$, they would demand E_2 workers. Point A (where W_0 and E_2 intersect) would lie on a new market demand curve, formed when demand shifted down because of the tax (remember, the wage on the vertical axis of Figure 3.4 is the wage *employees receive*, not the wage employers pay). Only if employee wages fell to $W_0 - X$ would firms want to continue hiring E_0 workers, for *employer* costs would then be the same as before the tax. Thus, point B would also be on the new, shifted demand curve. Note that with a tax of X , the new demand curve (D_1) is parallel to the old one, and the vertical distance between the two is X .

Now, the tax-related shift in the market demand curve to D_1 implies that there would be an excess supply of labor at the previous equilibrium wage of W_0 . This surplus of labor would create downward pressure on the *employee* wage, and this downward pressure would continue to be exerted until the employee wage fell to W_1 , the point at which the quantity of labor supplied just equaled the quantity demanded. At this point, employment would have also fallen to E_1 . Thus, *employees* bear a burden in the form of *lower wage rates and lower employment levels*. The lesson is clear: *employees* are not exempted from bearing costs

Figure 5.4

The Market Demand Curve and Effects of an Employer-Financed Payroll Tax



when the government chooses to generate revenues through a payroll tax on employers.

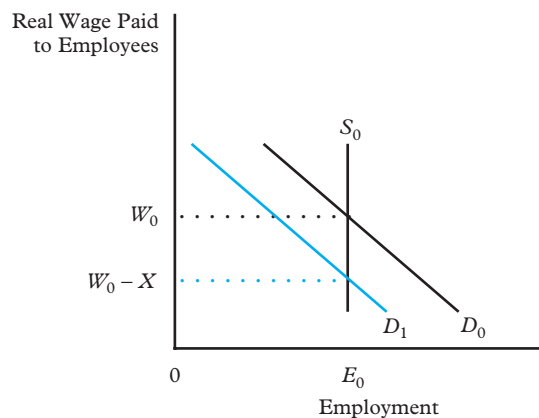
Figure 3.4 does suggest, however, that employers may bear at least *some* of the tax, because the wages received by employees do not fall by the full amount of the tax ($W_0 - W_1$ is smaller than X , which is the vertical distance between the two demand curves). This occurs because, with an upward-sloping labor market supply curve, employees withdraw labor as their wages fall, and it becomes more difficult for firms to find workers. If wages fell to $W_0 - X$, the withdrawal of workers would create a labor shortage that would drive wages to some point (W_1 in our example) between W_0 and $W_0 - X$. Only if the labor market supply curve were *vertical*—meaning that lower wages have no effect on labor supply—would the *entire amount of the tax* be shifted to workers in the form of a decrease in their wages by the amount of X (see Figure 3.5).

Effects of Labor Supply Curves The extent to which the labor market *supply* curve is sensitive to wages affects the proportion of the employer payroll tax that gets shifted to employees' wages. The less responsive labor supply is to changes in wages, the fewer the employees who withdraw from the market and the higher the proportion of the tax that gets shifted to workers in the form of a wage decrease (compare the outcomes in Figures 3.4 and 3.5). It must also be pointed out, however, that to the degree employee wages do *not* fall, employment levels *will*; when employee wages do not fall much in the face of an employer payroll-tax increase, employer labor costs are increased—and this increase reduces the quantity of labor employers demand.

A number of empirical studies have sought to ascertain what fraction of employers' payroll-tax costs are actually passed on to employees in the form of lower wages (or lower wage increases). Although the evidence is somewhat ambiguous, a comprehensive review of these studies led to at least a tentative

Figure 5.5

Payroll Tax with a Vertical Supply Curve



conclusion that most of a payroll tax is eventually shifted to wages, with little long-run effect on employment.⁹

Employment Subsidies as a Device to Help the Poor

The opposite of a payroll tax on employers is a government subsidy of employers' payrolls. In Figure 3.4, for example, if instead of *taxing* each hour of labor by X the government *paid* the employer X , the market labor demand curve would shift *upward* by a vertical distance of X . This upward movement of the demand curve would create pressures to increase employment and the wages received by employees; as with a payroll tax, whether the eventual effects would be felt more on employment or on wage rates depends on the shape of the labor market supply curve.

(Students should test their understanding in this area by drawing labor demand curves that reflect a new payroll subsidy of X per hour and then analyzing the effects on employment and employee wages with market supply curves that are, alternatively, upward-sloping and vertical. *Hint:* The outcomes should be those that would be obtained if demand curve D_1 in Figures 3.4 and 3.5 were shifted by the subsidy to curve D_0 .)

Payroll subsidies to employers can take many forms. They can be in the form of cash payments, as implied by the above hypothetical example, or they can be in the form of tax credits. These credits might directly reduce a firm's payroll-tax rate or they might reduce some other tax by an amount proportional to the number of labor hours hired; in either case, the credit has the effect of reducing the cost of hiring labor.

Furthermore, wage subsidies can apply to a firm's employment *level*, to any *new* employees hired after a certain date (even if they just replace workers who have left), or only to new hires that serve to *increase* the firm's level of employment. Finally, subsidies can be either *general* or *selective*. A general subsidy is not conditional on the characteristics of the people hired, whereas a selective, or *targeted*, plan makes the subsidy conditional on hiring people from certain target groups (such as the disadvantaged).

Experience in the United States with targeted wage subsidies has been modest. The Targeted Jobs Tax Credit (TJTC) program, which began in 1979 and was changed slightly over the years until it was finally discontinued in 1995, targeted disadvantaged youth, the handicapped, and welfare recipients, providing their employers with a tax credit that lasted for one year. In practice, the average duration

⁹Daniel S. Hamermesh, *Labor Demand* (Princeton, N.J.: Princeton University Press, 1993), 169–173. Also see Patricia M. Anderson and Bruce D. Meyer, "The Effects of the Unemployment Insurance Payroll Tax on Wages, Employment, Claims and Denials," *Journal of Public Economics* 78 (October 2000): 81–106; and Kevin Lang, "The Effect of the Payroll Tax on Earnings: A Test of Competing Models of Wage Determination," National Bureau of Economic Research Working Paper No. 9537, February 2003. Less wage and more job loss is reported in Adriana Kugler and Maurice Kugler, "Labor Market Effects of Payroll Taxes in Developing Countries: Evidence from Colombia," *Economic Development and Cultural Change* 57 (January 2009): 335–358.

EMPIRICAL STUDY

DO WOMEN PAY FOR EMPLOYER-FUNDED MATERNITY BENEFITS? USING CROSS-SECTION DATA OVER TIME TO ANALYZE “DIFFERENCES IN DIFFERENCES”

During the last half of 1976, Illinois, New Jersey, and New York passed laws requiring that employer-provided health insurance plans treat pregnancy the same as illness (that is, coverage of doctor’s bills and hospital costs had to be the same for pregnancy as for illnesses or injuries). These mandates increased the cost of health insurance for women of childbearing age by an amount that was equal to about 4 percent of their earnings. Were these increases in employer costs borne by employers or did they reduce the wages of women by an equivalent amount?

A problem confronting researchers on this topic is that the adopting states are all states with high incomes and likely to have state legislation encouraging the expansion of employment opportunities for women. Thus, comparing wage *levels* across states would require that we statistically control for all the factors, besides the maternity-benefit mandate, that affect wages. Because we can never be sure that we have adequate controls for the economic, social, and legal factors that affect wage levels by state, we need to find another way to perform the analysis.

Fortunately, answering the research question is facilitated by several factors: (a) some states adopted these laws and some did not; (b) even in states that adopted these laws, the insurance cost increases applied only to women (and

their husbands) of childbearing age and not to single men or older workers; and (c) the adopting states passed these laws during the same time period, so variables affecting women’s wages that change over time (such as recessions or the rising presence of women in the labor force) do not cloud the analysis.

Factors (a) and (c) above allow the conduct of what economists call a “differences-in-differences” analysis. Specifically, these factors allow us to compare wage *changes*, from the pre-adoption years to the post-adoption ones, among women of childbearing age in adopting states (the “experimental group”) to wage changes over the same period for women of the same age in states that did not adopt (a “comparison group”). By comparing within-state *changes* in wages, we avoid the need to find measures that would control for the economic, social, and public-policy forces that make the initial wage *level* in one state differ from that in another; whatever the factors are that raise wage levels in New Jersey, for example, they were there both before and after the adoption of mandated maternity benefits.

One might argue, of course, that the adopting and nonadopting states were subject to *other* forces (unrelated to maternity benefits) that led to different degrees of wage change over this period. For example, the economy of New Jersey might have been booming

during the period when maternity benefits were adopted, while economies elsewhere might not have been. However, if an adopting state is experiencing unique wage pressures in addition to those imposed by maternity benefits, the effects of these other pressures should show up in the wage changes experienced by single men or older women—groups in the adopting states that were not affected by the mandate. Thus, we can exploit factor (b) above by also comparing the wage changes for women of childbearing age in adopting states to those for single men or older women in the same states.

The three factors above enabled one researcher to measure how the wages of married women, aged 20–40, changed from 1974–1975 to 1977–1978 in the three adopting states. These changes were then

compared to changes in wages for married women of the same age in nonadopting (but economically similar) states. To account for forces *other than changing maternity benefits* that could affect wage changes across states during this period, the researcher also measured changes in wages for unmarried men and workers over 40 years of age. This study concluded that in the states adopting mandated maternity benefits, the post-adoption wages of women in the 20–40 age group were about 4 percent lower than they would have been without adoption. This finding suggests that the entire cost of maternity benefits was quickly shifted to women of childbearing age.

Source: Jonathan Gruber, “The Incidence of Mandated Maternity Benefits,” *American Economic Review* 84 (June 1994): 622–641.

of jobs under this program was six months, and the subsidy reduced employer wage costs by about 15 percent for jobs of this duration.

One problem that limited the effectiveness of the TJTC program was that the eligibility requirements for many of its participants were stigmatizing; that is, being eligible (on welfare, for example) was often seen by employers as a negative indicator of productivity. Nevertheless, one evaluation found that the employment of disadvantaged youth was enhanced by the TJTC. Specifically, it found that when 23- to 24-year-olds were removed from eligibility for the TJTC by changes in 1989, employment of disadvantaged youths of that age fell by over 7 percent.¹⁰ A more recent study found that the immediate employment and wage effects of a payroll subsidy were positive, but relatively small and not sustained.¹¹

¹⁰Lawrence F. Katz, “Wage Subsidies for the Disadvantaged,” in *Generating Jobs: How to Increase Demand for Less-Skilled Workers*, eds. Richard B. Freeman and Peter Gottschalk (New York: Russell Sage Foundation, 1998): 21–53.

¹¹Sasrah Hamersma, “The Effects of an Employer Subsidy on Employment Outcomes: A Study of the Work Opportunity and Welfare-to-Work Tax Credits,” *Journal of Policy Analysis and Management* 27 (Summer 2008): 498–520. A more positive view of the potential for payroll subsidies to increase employment can be found in Timothy J. Bartik and John H. Bishop, “The Job Creation Tax Credit,” Economic Policy Institute Briefing Paper No. 248 (Washington, D.C.: October 20, 2009).

Review Questions

1. In a statement during the 1992 presidential campaign, one organization attempting to influence the political parties argued that the wages paid by U.S. firms in their Mexican plants were so low that they “have no relationship with worker productivity.” Comment on this statement using the principles of profit maximization.
2. Assume that wages for keyboarders (data entry clerks) are lower in India than in the United States. Does this mean that keyboarding jobs in the United States will be lost to India? Explain.
3. The Occupational Safety and Health Administration promulgates safety and health standards. These standards typically apply to machinery (capital), which is required to be equipped with guards, shields, and the like. An alternative to these standards is to require the employer to furnish personal protective devices to employees (labor)—such as earplugs, hard hats, and safety shoes. *Disregarding* the issue of which alternative approach offers greater protection from injury, what aspects of each alternative must be taken into account when analyzing the possible *employment* effects of the two approaches to safety?
4. Suppose that prisons historically have required inmates to perform, *without pay*, various cleaning and food preparation jobs within the prison. Now, suppose that prisoners are offered paid work in factory jobs within the prison walls and that the cleaning and food preparation tasks are now performed by nonprisoners hired to do them. Would you expect to see any differences in the *technologies* used to perform these tasks? Explain.
5. Years ago, Great Britain adopted a program that placed a tax—to be collected from employers—on wages in *service* industries. Wages in manufacturing industries were not taxed. Discuss the wage and employment effects of this tax policy.
6. Suppose the government were to subsidize the wages of all women in the population by paying their *employers* 50 cents for every hour they work. What would be the effect on the wage rate women received? What would be the effect on the net wage employers paid? (The net wage would be the wage women received less 50 cents.)
7. In the last two decades, the United States has been subject to huge increases in the illegal immigration of workers from Mexico, most of them unskilled, and the government has considered ways to reduce the flow. One policy is to impose larger financial penalties on employers who are discovered to have hired illegal immigrants. What effect would this policy have on the employment of unskilled illegal immigrants? What effect would it have on the demand for skilled “native” labor?
8. If anti-sweatshop movements are successful in raising pay and improving working conditions for apparel workers in foreign countries, how will these changes abroad affect labor market outcomes for workers in the apparel and retailing industries in the United States? Explain.
9. The unemployment rate in France is currently over 10 percent, and the youth (under age 25) unemployment rate is about 22 percent. Over the next few years, one million people on the unemployment rolls will be offered subsidized jobs (the government subsidy will go to employers who create new jobs, and the subsidy will be X euros per hour per employee hired). Use the theory studied in this course to analyze how wage subsidies to employers are likely to affect employment levels in France.

Problems

1. An experiment conducted in Tennessee found that the scores of second graders and third graders on standardized tests for reading, math, listening, and word study skills were the same in small classrooms (13 to 17 students) as in regular classrooms (22 to 25 students). Suppose that there is a school that had 90 third graders taught by four teachers that added two additional teachers to reduce class sizes. If the Tennessee study can be generalized, what is the marginal product of labor (MP_L) of these two additional teachers?
2. The marginal revenue product of labor at the local sawmill is $MRP_L = 20 - 0.5L$, where L = the number of workers. If the wage of sawmill workers is \$10 per hour, then how many workers will the mill hire?
3. Suppose that the supply curve for lifeguards is $L_S = 20$, and the demand curve for lifeguards is $L_D = 100 - 20W$, where L = the number of lifeguards and W = the hourly wage. Graph both the demand and supply curves. Now, suppose that the government imposes a tax of \$1 per hour per worker on companies hiring lifeguards. Draw the new (after-tax) demand curve in terms of the employee wage. How will this tax affect the wage of lifeguards and the number employed as lifeguards?
4. The output of workers at a factory depends on the number of supervisors hired (see the following table). The factory sells its output for \$0.50 each, it hires 50 production workers at a wage of \$100 per day, and it needs to decide how many supervisors to hire. The daily wage of supervisors is \$500, but output rises as more supervisors are hired, as shown in the table. How many supervisors should it hire?

Supervisors	Output (Units per Day)
0	11,000
1	14,800
2	18,000
3	19,500
4	20,200
5	20,600

5. (Appendix) The Hormsbury Corporation produces yo-yos at its factory. Both its labor and capital markets are competitive. Wages are \$12 per hour, and yo-yo-making equipment (a computer-controlled plastic extruding machine) rents for \$4 per hour. The production function is $q = 40K^{0.25}L^{0.75}$, where q = boxes of yo-yos per week, K = hours of yo-yo equipment used, and L = hours of labor. Therefore, $MP_L = 30K^{0.25}L^{-0.25}$ and $MP_K = 10K^{-0.75}L^{0.75}$. Determine the cost-minimizing capital-labor ratio at this firm.
6. The following table shows the number of cakes that could be baked daily at a local bakery, depending on the number of bakers.

Number of Bakers	Number of Cakes
0	0
1	10
2	18
3	23
4	27

- a. Calculate the MP_L .
- b. Do you observe the law of diminishing marginal returns? Explain.
- c. Suppose each cake sells for \$10. Calculate the MRP_L .
- d. Draw the MRP_L curve, which is the demand curve for bakers.

- e. If each baker is paid \$80 per day, how many bakers will the bakery owner hire, given that the goal is to maximize profits? How many cakes will be baked and sold each day?
7. (Appendix) Creative Dangles is an earring design and manufacturing company. The production function for earrings is $Q = 25KL$, where Q = pairs of earrings per week, K = hours of equipment used, and L = hours of labor. Workers are paid \$8 per hour, and the equipment rents for \$8 per hour.
 - a. Determine the cost-minimizing capital-labor ratio at this firm.
 - b. How much does it cost to produce 10,000 pairs of earrings?
 - c. Suppose the rental cost of equipment decreases to \$6 per hour. What is the new cost-minimizing capital-labor ratio?
8. The demand curve for gardeners is $G_D = 19 - W$, where G = the number of gardeners, and W = the hourly wage. The supply curve is $G_S = 4 + 2W$.
 - a. Graph the demand curve and the supply curve. What is the equilibrium wage and equilibrium number of gardeners hired?
 - b. Suppose the town government imposes a \$2 per hour tax on all gardeners. Indicate the effect of the tax on the market for gardeners. What is the effect on the equilibrium wage and the equilibrium number of gardeners hired? How much does the gardener receive? How much does the customer pay? How much does the government receive as tax revenue?

Selected Readings

Blank, Rebecca M., ed. *Social Protection Versus Economic Flexibility: Is There a Trade-Off?* Chicago: University of Chicago Press, 1994.

Hamermesh, Daniel S. *Labor Demand*. Princeton, N.J.: Princeton University Press, 1993.

Katz, Lawrence F. "Wage Subsidies for the Disadvantaged." In *Generating Jobs: How to Increase Demand for Less-Skilled Workers*, eds. Richard B. Freeman and Peter Gottschalk, 21–53. New York: Russell Sage Foundation, 1998.

Graphical Derivation of a Firm's Labor Demand Curve

This chapter describes verbally the derivation of a firm's labor demand curve. This appendix will present the *same* derivation graphically. This graphical representation permits a more rigorous derivation, but our conclusion that demand curves slope downward in both the short and the long run will remain unchanged.

The Production Function

Output can generally be viewed as being produced by combining capital and labor. Figure 3A.1 illustrates this production function graphically and depicts several aspects of the production process.

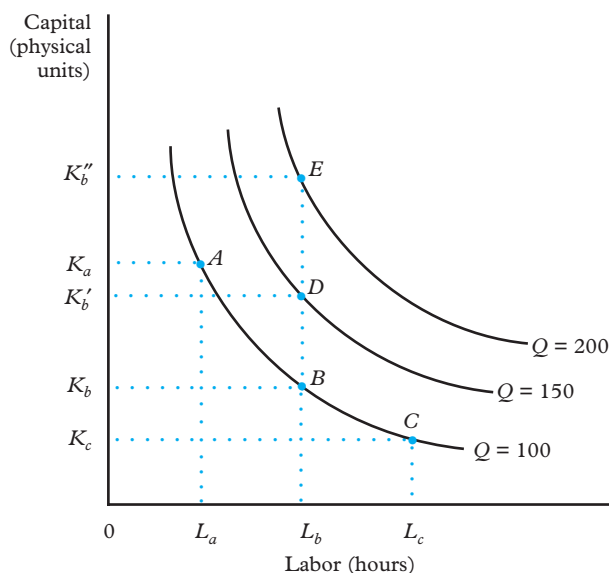
Consider the convex curve labeled $Q = 100$. Along this line, every combination of labor (L) and capital (K) produces 100 units of output (Q). That is, the combination of labor and capital at point A (L_a, K_a) generates the same 100 units of output as the combinations at points B and C . Because each point along the $Q = 100$ curve generates the same output, that curve is called an *isoquant* (*iso* = "equal"; *quant* = "quantity").

Two other isoquants are shown in Figure 3A.1 ($Q = 150, Q = 200$). These isoquants represent higher levels of output than the $Q = 100$ curve. The fact that these isoquants indicate higher output levels can be seen by holding labor constant at L_b (say) and then observing the different levels of capital. If L_b is combined with K_b in capital, 100 units of Q are produced. If L_b is combined with K'_b , 150 units are produced (K'_b is greater than K_b). If L_b is combined with even more capital (K''_b , say), 200 units of Q could be produced.

Note that the isoquants in Figure 3A.1 have *negative* slopes, reflecting an assumption that labor and capital are substitutes. If, for example, we cut capital from K_a to K_b , we could keep output constant (at 100) by increasing labor from L_a to L_b . Labor, in other words, could be substituted for capital to maintain a given production level.

Figure 5A.1

A Production Function



Finally, note the *convexity* of the isoquants. At point A, the $Q = 100$ isoquant has a steep slope, suggesting that to keep Q constant at 100, a given decrease in capital could be accompanied by a *modest* increase in labor. At point C, however, the slope of the isoquant is relatively flat. This flatter slope means that the same given decrease in capital would require a much *larger* increase in labor for output to be held constant. The decrease in capital permitted by a given increase in labor while output is being held constant is called the *marginal rate of technical substitution* (MRTS) between capital and labor. Symbolically, the MRTS can be written as

$$MRTS = \frac{\Delta K}{\Delta L} \Big|_{\bar{Q}} \quad (3.A1)$$

where Δ means “change in” and \bar{Q} means “holding output constant.” The MRTS is negative because if L is increased, K must be reduced to keep Q constant.

Why does the absolute value of the MRTS diminish as labor increases? When labor is highly used in the production process and capital is not very prevalent (point C in Figure 3A.1), there are many jobs that capital can do. Labor is easy to replace; if capital is increased, it will be used as a substitute for labor in parts of the production process where it will have the highest payoff. As capital becomes progressively more utilized and labor less so, the few remaining workers will be

doing jobs that are hardest for a machine to do, at which point it will take a lot of capital to substitute for a worker.¹

Demand for Labor in the Short Run

This chapter argues that firms will maximize profits in the short run (K fixed) by hiring labor until labor's marginal product (MP_L) is equal to the real wage (W/P). The reason for this decision rule is that the real wage represents the *cost* of an added unit of labor (in terms of output), while the marginal product is the *output* added by the extra unit of labor. As long as the firm, by increasing labor (K fixed), gains more in output than it loses in costs, it will continue to hire employees. The firm will stop hiring when the marginal cost of added labor exceeds MP_L .

The requirement that $MP_L = W/P$ in order for profits to be maximized means that the firm's labor demand curve in the short run (in terms of the *real* wage) is identical to its MP_L schedule (refer to Figure 3.1). Remembering that the MP_L is the extra output produced by one-unit increases in the amount of labor employed, holding capital constant, consider the production function displayed in Figure 3A.2. Holding capital constant at K_a , the firm can produce 100 units of Q if it employs labor equal to L_a . If labor is increased to L'_a , the firm can produce 50 more units of Q ; if labor is increased from L'_a to L''_a , the firm can produce an additional 50 units. Notice, however, that the required increase in labor to get the latter 50 units of added output, $L''_a - L'_a$, is larger than the extra labor required to produce the first 50-unit increment ($L'_a - L_a$). This difference can only mean that as labor is increased when K is held constant, each successive labor hour hired generates progressively smaller increments in output. Put differently, Figure 3A.2 graphically illustrates the diminishing marginal productivity of labor.

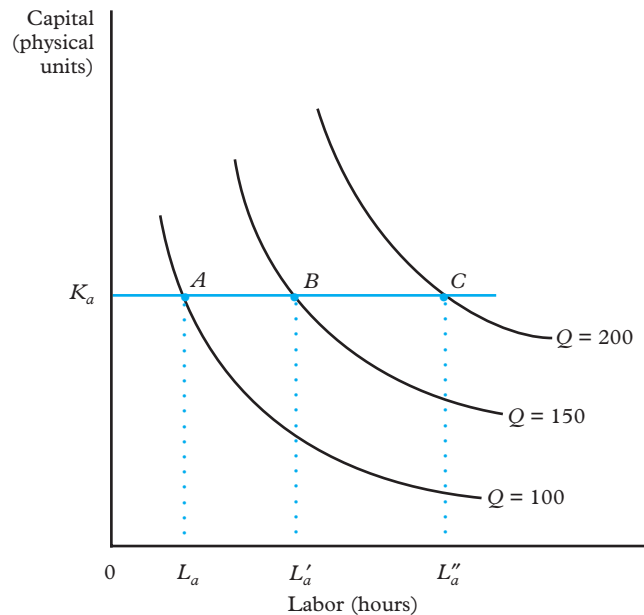
Why does labor's marginal productivity decline? This chapter explains that labor's marginal productivity declines because, with K fixed, each added worker has less capital (per capita) with which to work. Is this explanation proven in Figure 3A.2? The answer is, regrettably, no. Figure 3A.2 is drawn *assuming* diminishing marginal productivity. Renumbering the isoquants could produce a different set of marginal productivities. (To see this, change $Q = 150$ to $Q = 200$, and change $Q = 200$ to $Q = 500$. Labor's marginal productivity would then rise.) However, the logic that labor's marginal product must eventually fall as labor is increased, holding buildings, machines, and tools constant, is compelling. Further, as this chapter points out, even if MP_L rises initially, the firm will stop hiring labor only in the range where MP_L is declining; as long as MP_L is above W/P and *rising*, it will pay to continue hiring.

The assumptions that MP_L declines eventually and that firms hire until $MP_L = W/P$ are the bases for the assertion that a firm's short-run demand curve

¹Here is one example. Over time, telephone operators (who used to place long-distance calls) were replaced by a very capital-intensive direct-dialing system. Those operators who remain employed, however, perform tasks that are the most difficult for a machine to perform—handling collect calls, dispensing directory assistance, and acting as troubleshooters when problems arise.

Figure 5A.2

The Declining Marginal Productivity of Labor



for labor slopes downward. The graphical, more rigorous derivation of the demand curve in this appendix confirms and supports the verbal analysis in the chapter. However, it also emphasizes more clearly than a verbal analysis can that the downward-sloping nature of the short-run labor demand curve is based on an *assumption*—however reasonable—that MP_L declines as employment is increased.

Demand for Labor in the Long Run

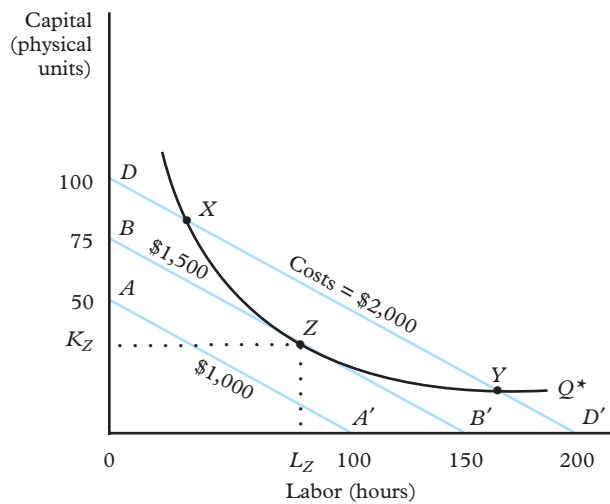
Recall that a firm maximizes its profits by producing at a level of output (Q^*) where marginal cost equals MR. That is, the firm will keep increasing output until the addition to its revenues generated by an extra unit of output just equals the marginal cost of producing that extra unit of output. Because MR, which is equal to output *price* for a competitive firm, is not shown in our graph of the production function, the profit-maximizing level of output cannot be determined. However, continuing our analysis of the production function can illustrate some important aspects of the demand for labor in the long run.

Conditions for Cost Minimization

In Figure 3A.3, profit-maximizing output is assumed to be Q^* . How will the firm combine labor and capital to produce Q^* ? It can maximize profits only if it produces Q^* in the least expensive way; that is, it must minimize the costs of

Figure 5A.3

Cost Minimization in the Production of Q^*
(Wage = \$10 per Hour; Price of a Unit of
Capital = \$20)



producing Q^* . To better understand the characteristics of cost minimization, refer to the three *isoexpenditure* lines— AA' , BB' , DD' —in Figure 3A.3. Along any one of these lines, the costs of employing labor and capital are equal.

For example, line AA' represents total costs of \$1,000. Given an hourly wage (W) of \$10 per hour, the firm could hire 100 hours of labor and incur total costs of \$1,000 if it used no capital (point A'). In contrast, if the price of a unit of capital (C) is \$20, the firm could produce at a total cost of \$1,000 by using 50 units of capital and no labor (point A). All the points between A and A' represent combinations of L and K that at $W = \$10$ and $C = \$20$, cost \$1,000 as well.

The problem with the isoexpenditure line of AA' is that it does not intersect the isoquant Q^* , implying that Q^* cannot be produced for \$1,000. At prices of $W = \$10$ and $C = \$20$, the firm cannot buy enough resources to produce output level Q^* and hold total costs to \$1,000. The firm can, however, produce Q^* for a total cost of \$2,000. Line DD' , representing expenditures of \$2,000, intersects the Q^* isoquant at points X and Y . The problem with these points, however, is that they are not cost-minimizing; Q^* can be produced for less than \$2,000.

Since isoquant Q^* is convex, the cost-minimizing combination of L and K in producing Q^* will come at a point where an isoexpenditure line is *tangent* to the isoquant (that is, just barely touches isoquant Q^* at only one place). Point Z , where labor equals L_Z and capital equals K_Z , is where Q^* can be produced at minimal cost, *given* that $W = \$10$ and $C = \$20$. No lower isoexpenditure curve touches the isoquant, meaning that Q^* cannot be produced for less than \$1,500.

An important characteristic of point Z is that the slope of the isoquant at point Z and the slope of the isoexpenditure line are the same (the slope of a curve at a given point is the slope of a line tangent to the curve at that point). The slope

of the isoquant at any given point is the *MRTS* as defined in equation (3A.1). Another way of expressing equation (3A.1) is

$$MRTS = \frac{-\Delta K/\Delta Q}{\Delta L/\Delta Q} \quad (3A.2)$$

Equation (3A.2) directly indicates that the *MRTS* is a ratio reflecting the reduction of capital required to decrease output by one unit if enough extra labor is hired so that output is tending to increase by one unit. (The ΔQ s in equation (3A.2) cancel each other and keep output constant.) Pursuing equation (3A.2) one step further, the numerator and denominator can be rearranged to obtain the following:²

$$MRTS = \frac{-\Delta K/\Delta Q}{\Delta L/\Delta Q} = \frac{-\Delta Q/\Delta L}{\Delta Q/\Delta K} = -\frac{MP_L}{MP_K} \quad (3A.3)$$

where MP_L and MP_K are the marginal productivities of labor and capital, respectively.

The slope of the *isoexpenditure line* is equal to the negative of the ratio W/C (in Figure 3A.3, W/C equals $10/20$, or 0.5).³ Thus, at point Z , where Q^* is produced in the minimum-cost fashion, the following equality holds:

$$MRTS = -\frac{MP_L}{MP_K} = -\frac{W}{C} \quad (3A.4)$$

Equation (3A.4) is simply a rearranged version of equation (3.8c).⁴

The economic meaning, or logic, behind the characteristics of cost minimization can most easily be seen by stating the *MRTS* as $-\frac{\Delta K/\Delta Q}{\Delta L/\Delta Q}$ (see equation 3A.2) and equating this version of the *MRTS* to $-\frac{W}{C}$:

$$\frac{\Delta K/\Delta Q}{\Delta L/\Delta Q} = -\frac{W}{C} \quad (3A.5)$$

or

$$\frac{\Delta K}{\Delta Q} \cdot C = \frac{\Delta L}{\Delta Q} \cdot W \quad (3A.6)$$

²This is done by making use of the fact that dividing one number by a second one is equivalent to multiplying the first by the *inverse* of the second.

³Note that $10/20 = 75/150$, or OB/OB' .

⁴The negative signs on each side of equation (3A.4) cancel each other and can therefore be ignored.

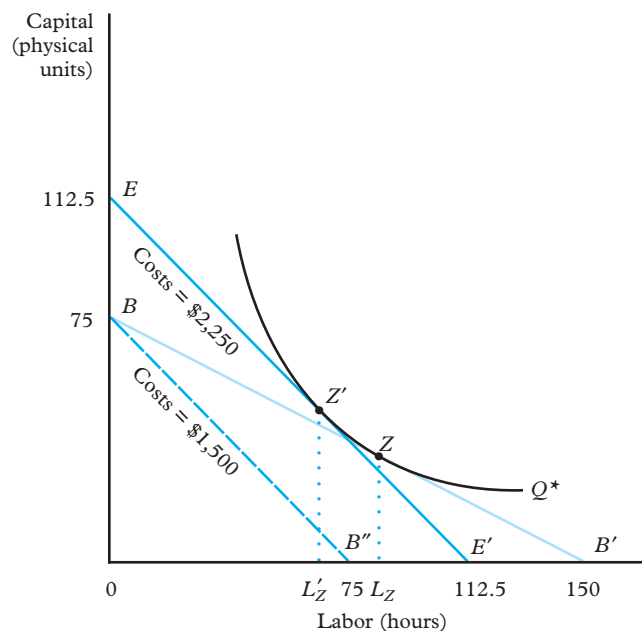
Equation (3A.6) makes it plain that to be minimizing costs, the cost of producing an extra unit of output by adding only labor must equal the cost of producing that extra unit by employing only additional capital. If these costs differed, the company could reduce total costs by expanding its use of the factor with which output can be increased more cheaply and cutting back on its use of the other factor. Any point where costs can still be reduced while Q is held constant is obviously not a point of cost minimization.

The Substitution Effect

If the wage rate, which was assumed to be \$10 per hour in Figure 3A.3, goes up to \$20 per hour (holding C constant), what will happen to the cost-minimizing way of producing output of Q^* ? Figure 3A.4 illustrates the answer that common sense would suggest: total costs rise, and more capital and less labor are used to produce Q^* . At $W = \$20$, 150 units of labor can no longer be purchased if total costs are to be held to \$1,500; in fact, if costs are to equal \$1,500, only 75 units of labor can be hired. Thus, the isoexpenditure curve for \$1,500 in costs shifts from BB' to BB'' and is no longer tangent to isoquant Q^* . Q^* can no longer be produced for \$1,500, and the cost of producing Q^* will rise. In Figure 3A.4, we assume the least-cost expenditure rises to \$2,250 (isoexpenditure line EE' is the one tangent to isoquant Q^*).

Figure 5A.4

Cost Minimization in the Production of Q^*
(Wage = \$20 per Hour; Price of a Unit of Capital = \$20)



Moreover, the increase in the cost of labor relative to capital induces the firm to use more capital and less labor. Graphically, the old tangency point of Z is replaced by a new one (Z'), where the marginal productivity of labor is higher relative to MP_K , as our discussions of equations (3.8c) and (3A.4) explained. Point Z' is reached (from Z) by adding more capital and reducing employment of labor. The movement from L_Z to L'_Z is the *substitution effect* generated by the wage increase.

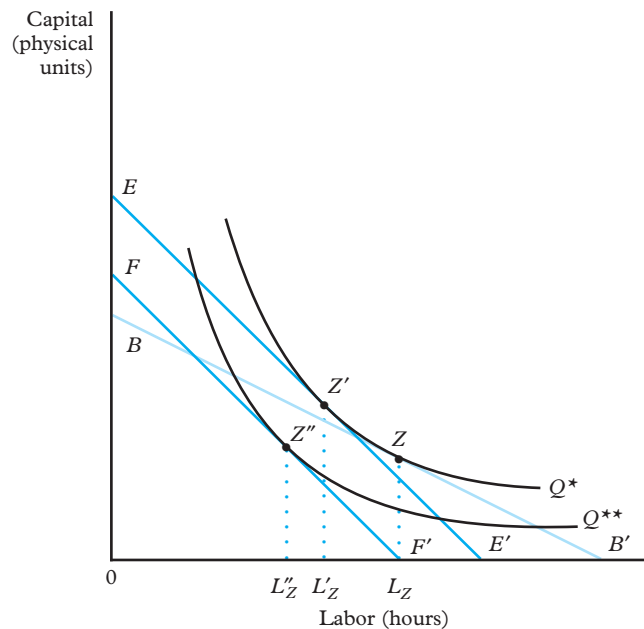
The Scale Effect

The fact that Q^* can no longer be produced for \$1,500, but instead involves at least \$2,250 in costs, will generally mean that it is no longer the profit-maximizing level of production. The new profit-maximizing level of production will be less than Q^* (how much less cannot be determined unless we know something about the product demand curve).

Suppose that the profit-maximizing level of output falls from Q^* to Q^{**} , as shown in Figure 3A.5. Since all isoexpenditure lines have the new slope of 21 when $W = \$20$ and $C = \$20$, the cost-minimizing way to produce Q^{**} will lie on an isoexpenditure line parallel to EE' . We find this cost-minimizing way to produce Q^{**} at point Z'' , where an isoexpenditure line (FF') is tangent to the Q^{**} isoquant.

Figure 5A.5

The Substitution and Scale Effects of a Wage Increase



The *overall* response in the employment of labor to an increase in the wage rate has been a fall in labor usage from L_z to L''_z . The decline from L_z to L'_z is called the substitution effect, as we have noted. It results because the *proportions* of K and L used in production change when the ratio of wages to capital prices (W/C) changes. The *scale effect* can be seen as the reduction in employment from L'_z to L''_z , wherein the usage of both K and L is cut back solely because of the reduced *scale* of production. Both effects are simultaneously present when wages increase and capital prices remain constant, but as Figure 3A.5 emphasizes, the effects are conceptually distinct and occur for different reasons. Together, these effects lead us to assert that the long-run labor demand curve slopes downward.

CHAPTER 6

Supply of Labor to the Economy: The Decision to Work

This and the next four chapters will focus on issues of *worker* behavior. That is, chapters 6–10 will discuss and analyze various aspects of *labor supply* behavior. Labor supply decisions can be roughly divided into two categories. The first, which is addressed in this chapter and the next, includes decisions about whether to work at all and, if so, how long to work. Questions that must be answered include whether to participate in the labor force, whether to seek part-time or full-time work, and how long to work both at home and for pay. The second category of decisions, which is addressed in chapters 8–10, deals with the questions that must be faced by a person who has decided to seek work for pay: the occupation or general class of occupations in which to seek offers (chapters 8 and 9) and the geographical area in which offers should be sought (chapter 10).

This chapter begins with some basic facts concerning labor force participation rates and hours of work. We then develop a theoretical framework that can be used in the analysis of decisions to work for pay. This framework is also useful for analyzing the structure of various income maintenance programs.

Trends in Labor Force Participation and Hours of Work

When a person actively seeks work, he or she is, by definition, in the *labor force*. As pointed out in chapter 2, the *labor force participation rate* is the percentage of a given population that either has a job or is looking for one. Thus, one clear-cut

statistic important in measuring people's willingness to work outside the home is the labor force participation rate.

Labor Force Participation Rates

One of the most dramatic changes in the labor market over the past six decades has been the increased labor force participation of women, especially married women. Table 6.1 shows the dimensions of this change. As recently as 1950, less than 25 percent of married women were in the labor force, but by 1980, this percentage had doubled. Recently, the labor force participation rate of married women has reached over 60 percent, although since 2000, the growth for married women seems to have stopped and the rates for single women have fallen.¹ One interest of this chapter is in understanding the forces underlying these changes.

Table 6.1

Labor Force Participation Rates of Females in the United States over 16 Years of Age, by Marital Status, 1900–2008 (Percentage)

Year	All Females	Single	Widowed, Divorced	Married
1900	20.6	45.9	32.5	5.6
1910	25.5	54.0	34.1	10.7
1920	24.0			9.0
1930	25.3	55.2	34.4	11.7
1940	26.7	53.1	33.7	13.8
1950	29.7	53.6	35.5	21.6
1960	37.7	58.6	41.6	31.9
1970	43.3	56.8	40.3	40.5
1980	51.5	64.4	43.6	49.8
1990	57.5	66.7	47.2	58.4
2000	59.9	68.9	49.0	61.1
2008	59.5	65.3	49.2	61.4

Sources: 1900–1950: Clarence D. Long, *The Labor Force under Changing Income and Employment* (Princeton, N.J.: Princeton University Press, 1958), Table A–6.

1960–2008: U.S. Department of Labor, Bureau of Labor Statistics, *Handbook of Labor Statistics*, Bulletin 2340 (Washington, D.C.: U.S. Government Printing Office, 1989), Table 6; and U.S. Census Bureau, 2010 *Statistical Abstract*, Section 12 (Table 583), <http://www.census.gov/compendia/statab/2010edition.html>.

¹Chinhui Juhn and Simon Potter, “Changes in Labor Force Participation in the United States,” *Journal of Economic Perspectives* 20 (Summer 2006): 27–46, offers a summary analysis of recent changes in the participation rates of both women and men.

As can be seen in Table 6.2, a second set of changes in labor force participation is the decrease in the participation rates of men, especially among the young and the old. The most substantial decreases in the United States have been among those 65 and older, from about 42 percent in 1950 to about half that currently—although since 1990 rates have been climbing a bit. Participation rates for men of “prime age” have declined only slightly since 1950, although among 45- to 64-year-olds, there were sharp decreases in the 1930s and 1970s. Clearly, men are starting their work lives later and ending them earlier than they were in 1950.

The trends in American labor force participation rates have also been observed in other industrialized countries. In Table 6.3, we display, for countries with comparable data, the trends in participation rates for women in the 25–54 age group and for men near the age of early retirement (55 to 64 years old). Typically, the fraction of women in the labor force rose from half or less in 1965 to three-quarters or more roughly 40 years later. Among men between the ages of 55 and 64, participation fell markedly in each country except Japan, although the declines were much larger in some countries (France, for example) than others

Table 6.2

Labor Force Participation Rates for Males in the United States, by Age, 1900–2008 (percentage)

Year	Age Groups					
	14–19	16–19	20–24	25–44	45–64	Over 65
1900	61.1		91.7	96.3	93.3	68.3
1910	56.2		91.1	96.6	93.6	58.1
1920	52.6		90.9	97.1	93.8	60.1
1930	41.1		89.9	97.5	94.1	58.3
1940	34.4		88.0	95.0	88.7	41.5
1950	39.9	63.2	82.8	92.8	87.9	41.6
1960	38.1	56.1	86.1	95.2	89.0	30.6
1970	35.8	56.1	80.9	94.4	87.3	25.0
1980		60.5	85.9	95.4	82.2	19.0
1990		55.7	84.4	94.8	80.5	16.3
2000		52.8	82.6	93.0	80.4	17.7
2008		40.1	78.7	91.9	81.4	21.5

Sources: 1900–1950: Clarence D. Long, *The Labor Force under Changing Income and Employment* (Princeton, N.J.: Princeton University Press, 1958), Table A–2.

1960: U.S. Department of Commerce, Bureau of the Census, *Census of Population, 1960: Employment Status*, Subject Reports PC(2)–6A, Table 1.

1970: U.S. Department of Commerce, Bureau of the Census, *Census of Population, 1970: Employment Status and Work Experience*, Subject Reports PC(2)–6A, Table 1.

1980–2008: U.S. Census Bureau, *2010 Statistical Abstract*, Section 12 (Table 575), <http://www.census.gov/compendia/statab/2010edition.html>.

Table 6.3**Labor Force Participation Rates of Women and Older Men, Selected Countries, 1965–2008 (Percentage)**

Country	1965	1973	1983	1993	2008
<i>Women, Age 25 to 54</i>					
Canada	33.9	44.0	65.1	75.6	82.0
France	42.8	54.1	67.0	76.1	83.2
Germany	46.1	50.5	58.3	72.5	80.5
Japan	—	53.0 ^a	59.5	65.2	70.3
Sweden	56.0	68.9	87.1	88.2	87.5
United States	45.1	52.0	67.1	74.6	75.8
<i>Men, Age 55 to 64</i>					
Canada	86.4	81.3	72.3	60.4	67.2
France	76.0	72.1	53.6	43.5	42.6
Germany	84.6	73.4	63.1	53.0	67.2
Japan	—	86.3 ^a	84.7	85.4	85.1
Sweden	88.3	82.7	77.0	70.9	76.7
United States	82.9	76.9	69.4	66.5	70.4

^aData are for 1974 (earlier data not comparable).

Source: Organisation for Economic Co-operation and Development, *Labour Force Statistics* (Paris: OECD, various dates).

(Sweden). Furthermore, the downward trends in four of the six countries shown appear to have reversed since the mid-1990s. Thus, while there are some differences in trends across the countries, it is likely that common forces are influencing labor supply trends in the industrialized world.

Hours of Work

Because data on labor force participation include both the employed and those who want a job but do not have one, they are a relatively pure measure of labor supply. In contrast, the weekly or yearly hours of work put in by the typical employee are often thought to be determined only by the demand side of the market. After all, don't employers, in responding to the factors discussed in chapter 5, set the hours of work expected of their employees? They do, of course, but hours worked are also influenced by *employee* preferences on the supply side of the market, especially in the long run.

Even though employers set work schedules, employees can exercise their preferences regarding hours of work through their choice of part-time or full-time work, their decisions to work at more than one job, or their selection of

occupations and employers.² For example, women managers who work full-time average more hours of work per week than full-time clerical workers, and male sales workers work more hours per week than their full-time counterparts in skilled craft jobs. Moreover, different employers offer different mixes of full-time and part-time work, require different weekly work schedules, and have different policies regarding vacations and paid holidays.

Employer offers regarding both hours and pay are intended to enhance their profits, but they must also satisfy the preferences of current and prospective employees. For example, if employees receiving an hourly wage of \$X for 40 hours per week really wanted to work only 30 hours at \$X per hour, some enterprising employer (presumably one with relatively lower quasi-fixed costs) would eventually seize on their dissatisfaction and offer jobs with a 30-hour workweek, ending up with a more satisfied, productive workforce in the process.

While the labor supply preferences of employees must be satisfied in the long run, most of the short-run changes in hours of work seem to emanate from the *demand* side of the market.³ Workweeks typically vary over the course of a business cycle, for example, with longer hours worked in periods of robust demand. In analyzing trends in hours of work, then, we must carefully distinguish between the forces of supply and demand.

In the first part of the twentieth century, workers in U.S. manufacturing plants typically worked 55 hours per week in years with strong economic activity; in the last two decades, American manufacturing workers have worked, on average, less than 40 hours per week during similar periods. For example, in the years 1988, 1995, and 2004—when the unemployment rate was roughly 5.5 percent and falling—manufacturing production workers averaged 38.4, 39.3, and 38.6 hours per week, respectively. In general, the decline in weekly hours of

²At any time, roughly 5 percent of American workers hold more than one job—although many more (20 percent of men and 12 percent of women) hold more than one job *at some point within a year*. See Christina H. Paxson and Nachum Sicherman, “The Dynamics of Dual Job Holding and Job Mobility,” *Journal of Labor Economics* 14 (July 1996): 357–393; and Jean Kimmel and Karen Smith Conway, “Who Moonlights and Why? Evidence from the SIPP,” *Industrial Relations* 40 (January 2001): 89–120. For a study that tests (and finds support for) the assumption that workers are *not* restricted in their choice of work hours, see John C. Ham and Kevin T. Reilly, “Testing Intertemporal Substitution, Implicit Contracts, and Hours Restriction Models of the Labor Market Using Micro Data,” *American Economic Review* 92 (September 2002): 905–927.

³See, for example, Joseph G. Altonji and Christina H. Paxson, “Job Characteristics and Hours of Work,” in *Research in Labor Economics*, vol. 8, ed. Ronald Ehrenberg (Greenwich, Conn.: JAI Press, 1986); Orley Ashenfelter, “Macroeconomic Analyses and Microeconomic Analyses of Labor Supply,” *Carnegie-Rochester Conference Series on Public Policy* 21 (1984): 117–156. A recent study has shown that workers’ desired labor supply adjustments come more from changing jobs than from changing hours with the same employer; see Richard Blundell, Mike Brewer, and Marco Francesconi, “Job Changes and Hours Changes: Understanding the Path of Labor Supply Adjustment,” *Journal of Labor Economics* 26 (July 2008): 421–454.

manufacturing work in the United States occurred prior to 1950, and since then, hours of work have shown little tendency to decline.⁴

A Theory of the Decision to Work

Can labor supply theory help us to understand the long-run trends in labor force participation and hours of work noted above? Because labor is the most abundant factor of production, it is fair to say that any country's well-being in the long run depends heavily on the willingness of its people to work. Leisure and other ways of spending time that do not involve work for pay are also important in generating well-being; however, any economy relies heavily on goods and services produced for market transactions. Therefore, it is important to understand the *work-incentive* effects of higher wages and incomes, different kinds of taxes, and various forms of income maintenance programs.

The decision to work is ultimately a decision about how to spend time. One way to use our available time is to spend it in pleasurable leisure activities. The other major way in which people use time is to work. We can work around the home, performing such *household production* as raising children, sewing, building, or even growing food. Alternatively, we can work for pay and use our earnings to purchase food, shelter, clothing, and child care.

Because working for pay and engaging in household production are two ways of getting the same jobs done, we shall initially ignore the distinction between them and treat work activities as working for pay. We shall therefore be characterizing the decision to work as a choice between leisure and working for pay. Most of the crucial factors affecting work incentives can be understood in this context, but insight into labor supply behavior can also be enriched by a consideration of household production; this we do in chapter 7.

If we regard the time spent eating, sleeping, and otherwise maintaining ourselves as more or less fixed by natural laws, then the discretionary time we have (16 hours a day, say) can be allocated to either work or leisure. It is most convenient for us to begin our analysis of the work/leisure choice by analyzing the *demand for leisure hours*.

Some Basic Concepts

Basically, the demand for a good is a function of three factors:

1. The *opportunity cost* of the good (which is often equal to *market price*).

⁴The averages cited in this paragraph refer to *actual* hours of work (obtained from the *Census of Manufactures*), not the more commonly available "hours paid for," which include paid time off for illness, holidays, and vacations. A recent study found an unexpected *expansion* of work hours among highly educated men during the last two decades of the twentieth century; see Peter Kuhn and Fernando Lozano, "The Expanding Workweek? Understanding Trends in Long Work Hours among U.S. Men, 1979–2006," *Journal of Labor Economics* 26 (April 2008): 311–343.

2. One's level of *wealth*.
3. One's set of *preferences*.

For example, consumption of heating oil will vary with the *cost* of such oil; as that cost rises, consumption tends to fall unless one of the other two factors intervenes. As *wealth* rises, people generally want larger and warmer houses that obviously require more oil to heat.⁵ Even if the price of energy and the level of personal wealth were to remain constant, the demand for energy could rise if a falling birthrate and lengthened life span resulted in a higher proportion of the population being aged and therefore wanting warmer houses. This change in the composition of the population amounts to a shift in the overall *preferences* for warmer houses and thus leads to a change in the demand for heating oil. (Economists usually assume that preferences are given and not subject to immediate change. For policy purposes, changes in prices and wealth are of paramount importance in explaining changes in demand because these variables are more susceptible to change by government or market forces.)

Opportunity Cost of Leisure To apply this general analysis of demand to the demand for leisure, we must first ask, "What is the opportunity cost of leisure?" The cost of spending an hour watching television is basically what one could earn if one had spent that hour working. Thus, the opportunity cost of an hour of leisure is equal to one's *wage rate*—the *extra earnings* a worker can take home from an *extra hour of work*.⁶

Wealth and Income Next, we must understand and be able to measure wealth. Naturally, wealth includes a family's holdings of bank accounts, financial investments, and physical property. Workers' skills can also be considered assets, since these skills can be, in effect, rented out to employers for a price. The more one can get in wages, the larger the value of one's human assets. Unfortunately, it is not usually possible to directly measure people's wealth. It is much easier to measure the *returns* from that wealth, because data on total *income* are readily available from government surveys. Economists often use total income as an indicator of total wealth, since the two are conceptually so closely related.⁷

Defining the Income Effect Theory suggests that if income increases while wages and preferences are held constant, the number of leisure hours demanded will rise. Put differently, *if income increases, holding wages constant, desired hours of*

⁵When the demand for a good rises with wealth, economists say the good is a *normal good*. If demand falls as wealth rises, the good is said to be an *inferior good* (traveling or commuting by bus is sometimes cited as an example of an inferior good).

⁶This assumes that individuals can work as many hours as they want at a fixed wage rate. While this assumption may seem overly simplistic, it will not lead to wrong conclusions with respect to the issues analyzed in this chapter. More rigorously, it should be said that leisure's marginal opportunity cost is the marginal wage rate (the wage one could receive for an extra hour of work).

⁷The best indicator of wealth is permanent, or long-run potential, income. Current income may differ from permanent income for a variety of reasons (unemployment, illness, unusually large amounts of overtime work, etc.). For our purposes here, however, the distinction between current income and permanent income is not too important.

work will go down. (Conversely, if income is reduced while the wage rate is held constant, desired hours of work will go up.) Economists call the response of desired hours of leisure to changes in income, with wages held constant, the *income effect*. The income effect is based on the simple notion that as incomes rise, holding leisure's opportunity cost constant, people will want to consume more leisure (which means working less).

Because we have assumed that time is spent either in leisure or in working for pay, the income effect can be expressed in terms of the *supply of working hours* as well as the demand for leisure hours. Because the ultimate focus of this chapter is labor supply, we choose to express this effect in the context of supply.

Using algebraic notation, we define the income effect as the change in hours of work (ΔH) produced by a change in income (ΔY), holding wages constant (\bar{W}):

$$\text{Income Effect} = \frac{\Delta H}{\Delta Y} \Big| \bar{W} < 0 \quad (6.1)$$

We say the income effect is *negative* because the *sign* of the *fraction* in equation (6.1) is *negative*. If income goes up (wages held constant), hours of work fall. If income goes down, hours of work increase. The numerator (ΔH) and denominator (ΔY) in equation (6.1) move in opposite directions, giving a negative sign to the income effect.

Defining the Substitution Effect Theory also suggests that *if income is held constant, an increase in the wage rate will raise the price and reduce the demand for leisure, thereby increasing work incentives.* (Likewise, a decrease in the wage rate will reduce leisure's opportunity cost and the incentives to work, holding income constant.) This *substitution effect* occurs because as the cost of leisure changes, income held constant, leisure and work hours are substituted for each other.

In contrast to the income effect, the substitution effect is *positive*. Because this effect is the change in hours of work (ΔH) induced by a change in the wage (ΔW), holding income constant (\bar{Y}), the substitution effect can be written as

$$\text{Substitution Effect} = \frac{\Delta H}{\Delta W} \Big| \bar{Y} > 0 \quad (6.2)$$

Because the numerator (ΔH) and denominator (ΔW) always move in the same direction, at least in theory, the substitution effect has a positive sign.

Observing Income and Substitution Effects Separately At times, it is possible to observe situations or programs that create only one effect or the other. (Laboratory experiments can also create separate income and substitution effects; an experiment with pigeons, discussed in Example 6.1, suggests that labor supply theory can even be generalized beyond humans!) Usually, however, both effects are simultaneously present, often working against each other.

EXAMPLE 6.1**The Labor Supply of Pigeons**

Economics has been defined as “the study of the allocation of scarce resources among unlimited and competing uses.” Stated this way, the tools of economics can be used to analyze the behavior of animals as well as humans. In a classic study, Raymond Battalio, Leonard Green, and John Kagel describe an experiment in which they estimated income and substitution effects (and thus the shape of the labor supply curve) for animals.

The subjects were male White Carneaux pigeons. The job task consisted of pecking at a response key. If the pigeons pecked the lever enough times, their payoff was access to a food hopper containing mixed grains. “Wages” were changed by altering the average number of pecks per payoff. Pecking requirements varied from as much as 400 pecks per payoff (a very low wage) to as few as 12.5 pecks. In addition, “unearned income” could be changed by giving the pigeons free access to the food hopper without the need for pecking. The environment was meant to observe the trade-off

between key pecking (“work”) and the pigeons’ primary alternative activities of preening themselves and walking around (“leisure”). The job task was not awkward or difficult for pigeons to perform, but it did require effort.

Battalio, Green, and Kagel found that pigeons’ actions were perfectly consistent with economic theory. In the first stage of the experiment, they cut the wage rate (payoff per peck) but added enough free food to isolate the substitution effect. In almost every case, the birds reduced their labor supply and spent more time on leisure activities. In the second stage of the experiment, they took away the free food to isolate the income effect. They found that every pigeon increased its pecking (cutting its leisure) as its income was cut. Thus, leisure is a normal good for pigeons.

Data from: Raymond C. Battalio, Leonard Green, and John H. Kagel, “Income-Leisure Tradeoffs of Animal Workers,” American Economic Review 71 (September 1981): 621–632.

Receiving an inheritance offers an example of the income effect by itself. The bequest enhances wealth (income) *independent* of the hours of work. Thus, income is increased *without* a change in the compensation received from an hour of work. In this case, the income effect induces the person to consume more leisure, thereby reducing the willingness to work. (Some support for this theoretical prediction can be seen later in Example 6.3.)

Observing the substitution effect by itself is rare, but one example comes from the 1980 presidential campaign, when candidate John Anderson proposed a program aimed at conserving gasoline. His plan consisted of raising the gasoline tax but offsetting this increase by a reduced Social Security tax payable by individuals on their earnings. The idea was to raise the price of gasoline without reducing people’s overall spendable income.

For our purposes, this plan is interesting because, for the typical worker, it would have created only a substitution effect on labor supply. Social Security revenues are collected by a tax on earnings, so reductions in the tax are, in effect, increases in the wage rate for most workers. For the average person, however, the increased wealth associated with this wage increase would have been exactly

offset by increases in the gasoline tax.⁸ Hence, wages would have been increased while income was held more or less constant. This program would have created a substitution effect that induced people to work more hours.

Both Effects Occur When Wages Rise While the above examples illustrate situations in which the income or the substitution effect is present by itself, *normally both effects are present, often working in opposite directions*. The presence of both effects working in opposite directions creates ambiguity in predicting the overall labor supply response in many cases. Consider the case of a person who receives a wage increase.

The labor supply response to a simple wage increase will involve *both* an income effect and a substitution effect. The *income effect* is the result of the worker's enhanced wealth (or potential income) after the increase. For a given level of work effort, he or she now has a greater command over resources than before (because more income is received for any given number of hours of work). The *substitution effect* results from the fact that the wage increase raises the opportunity costs of leisure. Because the actual labor supply response is the *sum* of the income and substitution effects, we cannot predict the response in advance; theory simply does not tell us which effect is stronger.

If the *income effect* is stronger, the person will respond to a wage increase by decreasing his or her labor supply. This decrease will be *smaller* than if the same change in wealth were due to an increase in *nonlabor* wealth, because the substitution effect is present and acts as a moderating influence. However, as seen in Example 6.2, when the *income effect* dominates, the substitution effect is not large enough to prevent labor supply from *declining*. It is entirely plausible, of course, that the *substitution effect* will dominate. If so, the actual response to wage increases will be to *increase* labor supply.

Should the substitution effect dominate, the person's labor supply curve—relating, say, his or her desired hours of work to wages—will be *positively sloped*. That is, labor supplied will increase with the wage rate. If, on the other hand, the income effect dominates, the person's labor supply curve will be *negatively sloped*. Economic theory cannot say which effect will dominate, and in fact, individual labor supply curves could be positively sloped in some ranges of the wage and negatively sloped in others. In Figure 6.1, for example, the person's desired hours of work increase (substitution effect dominates) when wages go up as long as wages are low (below W^*). At higher wages, however, further increases result in

⁸An increase in the price of gasoline will reduce the income people have left for expenditures on non-gasoline consumption only if the demand for gasoline is inelastic. In this case, the percentage reduction in gasoline consumption is smaller than the percentage increase in price; total expenditures on gasoline would thus rise. Our analysis assumes this to be the case. For a study of how gasoline taxes affect labor supply, see Sarah West and Robertson Williams, "Empirical Estimates for Environmental Policy Making in a Second-Best Setting," National Bureau of Economic Research, Working Paper No. 10330 (March 2004).

EXAMPLE 6.2**The Labor Supply of New York City Taxi Drivers**

Testing the theory of labor supply is made difficult by the fact that most workers cannot change their hours of work very much without changing jobs. Taxi drivers in New York City, however, do choose their own hours of work, so it is interesting to see how their hours of work—reflected in miles driven—responded to fare increases approved by the city's Taxi and Limousine Commission in 1996 and 2004. These fare increases raised the hourly

pay of taxi drivers by an average of 19 percent, and a careful study of how drivers responded found that they *reduced* their miles driven by about 4 percent. Clearly, then, the income effect of their wage increases was stronger than the substitution effect.

Source: Orley Ashenfelter, Kirk Doran, and Bruce Schaller, "A Shred of Credible Evidence on the Long-Run Elasticity of Labor Supply," *Economica* 77 (October, 2010): 637–650.

reduced hours of work (the income effect dominates); economists refer to such a curve as *backward-bending*.

Analysis of the Labor/Leisure Choice

This section introduces indifference curves and budget constraints—visual aids that make the theory of labor supply easier to understand and to apply to complex policy issues. These graphical aids visually depict the basic factors underlying the demand for leisure (supply of labor) discussed earlier.

Preferences Let us assume that there are two major categories of goods that make people happy—leisure time and the goods people can buy with money. If we take the prices of goods as fixed, then they can be compressed into one index that is measured by money income (with prices fixed, more money income means

Figure 6.1

An Individual Labor Supply Curve Can Bend Backward

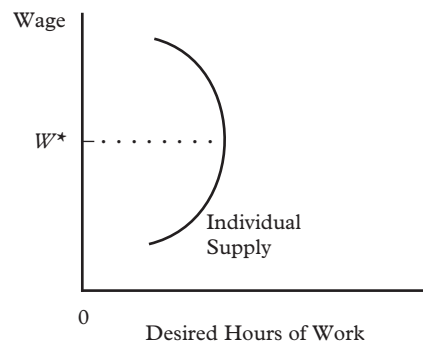
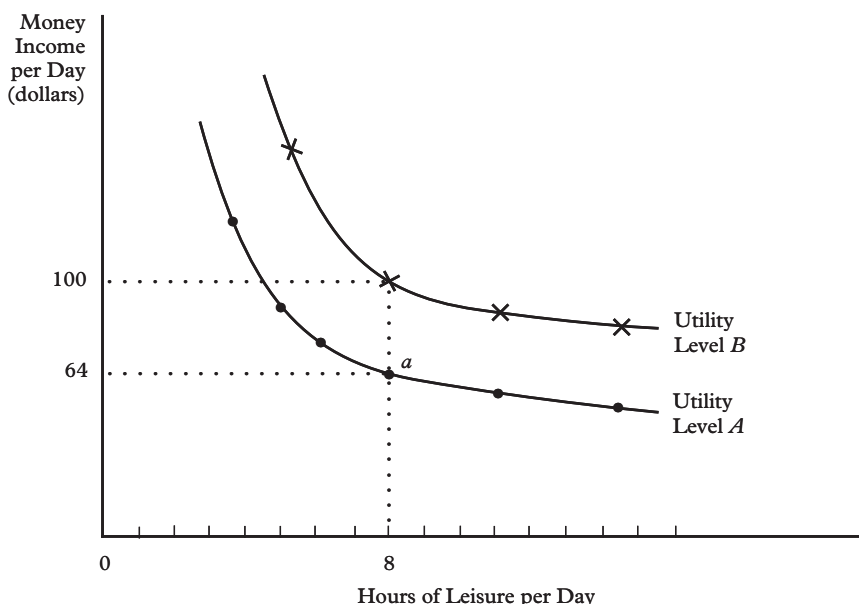


Figure 6.2

Two Indifference
Curves for the Same
Person



it is possible to consume more goods). Using two categories, leisure and money income, allows our graphs to be drawn in two-dimensional space.

Since both leisure and money can be used to generate satisfaction (or *utility*), these two goods are to some extent substitutes for each other. If forced to give up some money income—by cutting back on hours of work, for example—some increase in leisure time could be substituted for this lost income to keep a person as happy as before.

To understand how preferences can be graphed, suppose a thoughtful consumer/worker were asked to decide how happy he or she would be with a daily income of \$64 combined with 8 hours of leisure (point *a* in Figure 6.2). This level of happiness could be called utility level *A*. Our consumer/worker could name *other combinations* of money income and leisure hours that would *also* yield utility level *A*. Assume that our respondent named five other combinations. All six combinations of money income and leisure hours that yield utility level *A* are represented by heavy dots in Figure 6.2. The curve connecting these dots is called an *indifference curve*, which connects the various combinations of money income and leisure that yield equal utility. (The term *indifference curve* is derived from the fact that since each point on the curve yields equal utility, a person is truly indifferent about where on the curve he or she will be.)

Our worker/consumer could no doubt achieve a higher level of happiness if he or she could combine the 8 hours of leisure with an income of \$100 per day

instead of just \$64 a day. This higher satisfaction level could be called utility level *B*. The consumer could name other combinations of money income and leisure that would also yield *this* higher level of utility. These combinations are denoted by the *Xs* in Figure 6.2 that are connected by a second indifference curve.

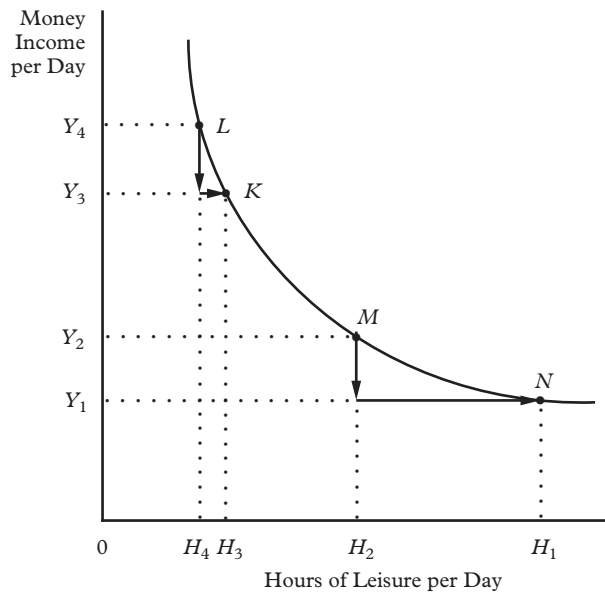
Indifference curves have certain specific characteristics that are reflected in the way they are drawn:

1. Utility level *B* represents more happiness than level *A*. Every level of leisure consumption is combined with a higher income on *B* than on *A*. Hence, our respondent prefers all points on indifference curve *B* to any point on curve *A*. A whole *set* of indifference curves could be drawn for this one person, each representing a different utility level. Any such curve that lies to the northeast of another one is preferred to any curve to the southwest because the northeastern curve represents a higher level of utility.
2. Indifference curves *do not intersect*. If they did, the point of intersection would represent one combination of money income and leisure that yielded two different levels of satisfaction. We assume our worker/consumer is not so inconsistent in stating his or her preferences that this could happen.
3. Indifference curves are *negatively sloped* because if either income or leisure hours are increased, the other is reduced in order to preserve the same level of utility. If the slope is steep, as at segment *LK* in Figure 6.3, a given loss of income need not be accompanied by a large increase in leisure hours to keep utility constant.⁹ When the curve is relatively flat, however, as at segment *MN* in Figure 6.3, a given decrease in income must be accompanied by a large increase in the consumption of leisure to hold utility constant. Thus, when indifference curves are relatively steep, people do not value money income as highly as when such curves are relatively flat; when they are flat, a loss of income can only be compensated for by a large increase in leisure if utility is to be kept constant.
4. Indifference curves are *convex*—steeper at the left than at the right. This shape reflects the assumption that when money income is relatively high and leisure hours are relatively few, leisure is more highly valued (and income less valued) than when leisure is abundant and income relatively scarce. At segment *LK* in Figure 6.3, a great loss of income (from Y_4 to Y_3 , for example) can be compensated for by just a little increase in leisure, whereas a little loss of leisure time (from H_3 to H_4 , for example) would require a relatively large increase in income to maintain equal utility. What is relatively scarce is more highly valued.

⁹Economists call the change in money income needed to hold utility constant when leisure hours are changed by one unit the *marginal rate of substitution* between leisure and money income. This marginal rate of substitution can be graphically understood as the slope of the indifference curve at any point. At point *L*, for example, the slope is relatively steep, so economists would say that the marginal rate of substitution at point *L* is relatively high.

Figure 6.3

An Indifference Curve

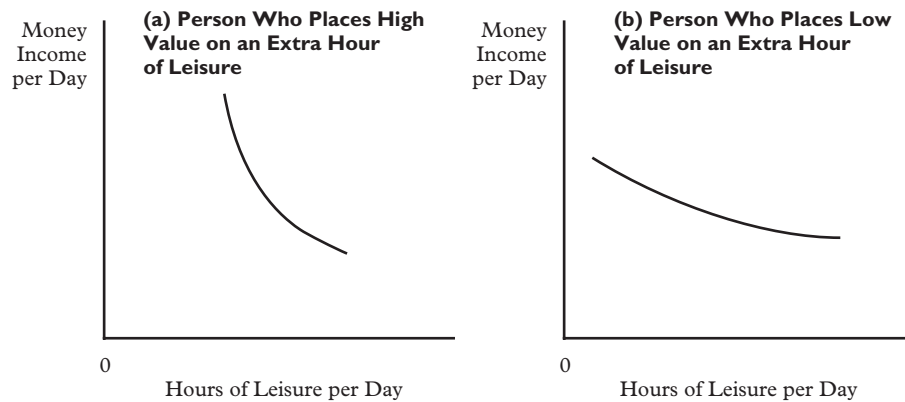


5. Conversely, when income is low and leisure is abundant (segment MN in Figure 6.3), income is more highly valued. Losing income (by moving from Y_2 to Y_1 , for example) would require a huge increase in leisure for utility to remain constant. To repeat, what is relatively scarce is assumed to be more highly valued.
6. Finally, different people have different sets of indifference curves. The curves drawn in Figures 6.2 and 6.3 were for *one person*. Another person would have a completely different set of curves. People who value leisure more highly, for example, would have had indifference curves that were generally steeper (see Figure 6.4a). People who do not value leisure highly would have relatively flat curves (see Figure 6.4b). Thus, individual preferences can be portrayed graphically.

Income and Wage Constraints Everyone would like to maximize his or her utility, which would be ideally done by consuming every available hour of leisure combined with the highest conceivable income. Unfortunately, the resources anyone can command are limited. Thus, all that is possible is to do the best one can, given limited resources. To see these resource limitations graphically requires superimposing constraints on one's set of indifference curves to see which combinations of income and leisure are available and which are not.

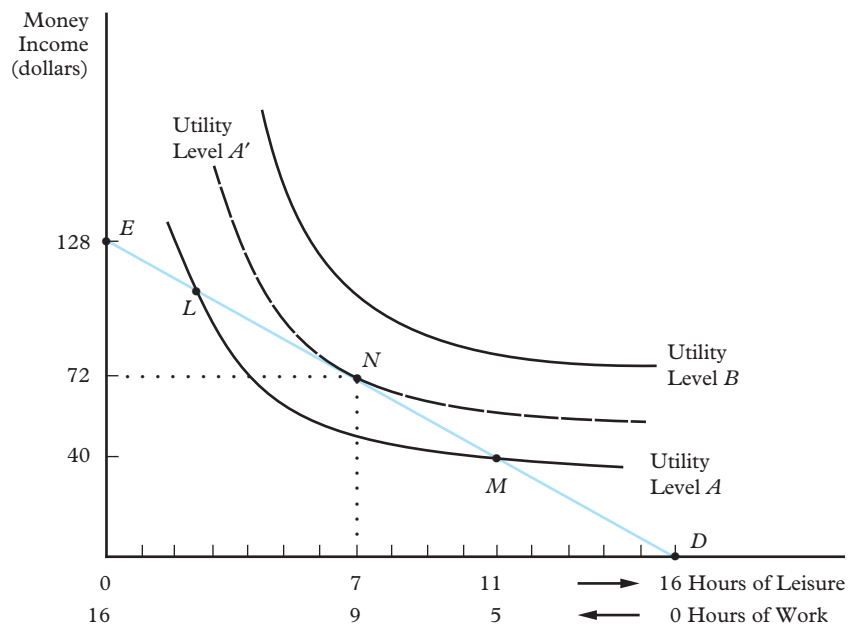
Suppose the person whose indifference curves are graphed in Figure 6.2 had no source of income other than labor earnings. Suppose, further, that he

Figure 6.4

Indifference
Curves for Two
Different People

or she could earn \$8 per hour. Figure 6.5 includes the two indifference curves shown in Figure 6.2 as well as a straight line (ED) connecting combinations of leisure and income that are possible for a person with an \$8 wage and no outside income. If 16 hours per day are available for work

Figure 6.5

Indifference Curves and
Budget Constraint

and leisure,¹⁰ and if this person consumes all 16 in leisure, then money income will be zero (point *D* in Figure 6.5). If 5 hours a day are devoted to work, income will be \$40 per day (point *M*), and if 16 hours a day are worked, income will be \$128 per day (point *E*). Other points on this line—for example, the point of 15 hours of leisure (1 hour of work) and \$8 of income—are also possible. This line, which reflects the combinations of leisure and income that are possible for the individual, is called the *budget constraint*. Any combination to the right of the budget constraint is not achievable; the person's command over resources is simply not sufficient to attain these combinations of leisure and money income.

The *slope* of the budget constraint is a graphical representation of the wage rate. One's wage rate is properly defined as the increment in income (ΔY) derived from an increment in hours of work (ΔH):

$$\text{Wage Rate} = \frac{\Delta Y}{\Delta H} \quad (6.3)$$

Now $\Delta Y/\Delta H$ is exactly the slope of the budget constraint (in absolute value).¹¹ Figure 6.5 shows how the constraint rises \$8 for every 1-hour increase in work: if the person works 0 hours, income per day is zero; if the person works 1 hour, \$8 in income is received; if he or she works 5 hours, \$40 in income is achieved. The constraint rises \$8 because the wage rate is \$8 per hour. If the person could earn \$16 per hour, the constraint would rise twice as fast and therefore be twice as steep.

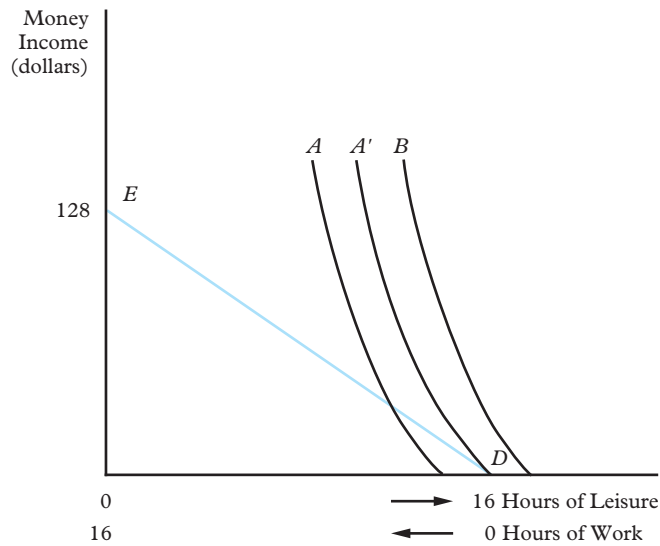
It is clear from Figure 6.5 that our consumer/worker cannot achieve utility level *B*. He or she can achieve *some* points on the indifference curve representing utility level *A*—specifically, those points between *L* and *M* in Figure 6.5. However, if our consumer/worker is a utility maximizer, he or she will realize that a utility level *above* *A* is possible. Remembering that an infinite number of indifference curves can be drawn between curves *A* and *B* in Figure 6.5, one representing each possible level of satisfaction between *A* and *B*, we can draw a curve (*A'*) that is northeast of curve *A* and just *tangent* to the budget constraint at point *N*. Any movement along the budget constraint *away* from the tangency point places the person on an indifference curve lying *below* *A'*.

¹⁰Our assumption that 8 hours per day are required for sleeping and other “maintenance” activities is purely for ease of exposition. These activities themselves are a matter of economic choice, at least to some extent; see, for example, Jeff E. Biddle and Daniel S. Hamermesh, “Sleep and the Allocation of Time,” *Journal of Political Economy* 98, no. 5, pt. 1 (October 1990): 922–943. Modeling a three-way choice between work, leisure, and maintenance activities would complicate our analysis without changing the essential insights theory can offer about the labor/leisure choice workers must make.

¹¹The vertical change for a one-unit change in horizontal distance is the definition of *slope*. *Absolute value* refers to the magnitude of the slope, disregarding whether it is positive or negative. The budget constraint drawn in Figure 6.5 is a straight line (and thus has a constant slope). In economic terms, a straight-line budget constraint reflects the assumption that the wage rate at which one can work is fixed and that it does not change with the hours of work. However, the major theoretical implications derived from using a straight-line constraint would be unchanged by employing a convex one, so we are using the fixed-wage assumption for ease of exposition.

Figure 6.6

The Decision Not to Work Is a
"Corner Solution"



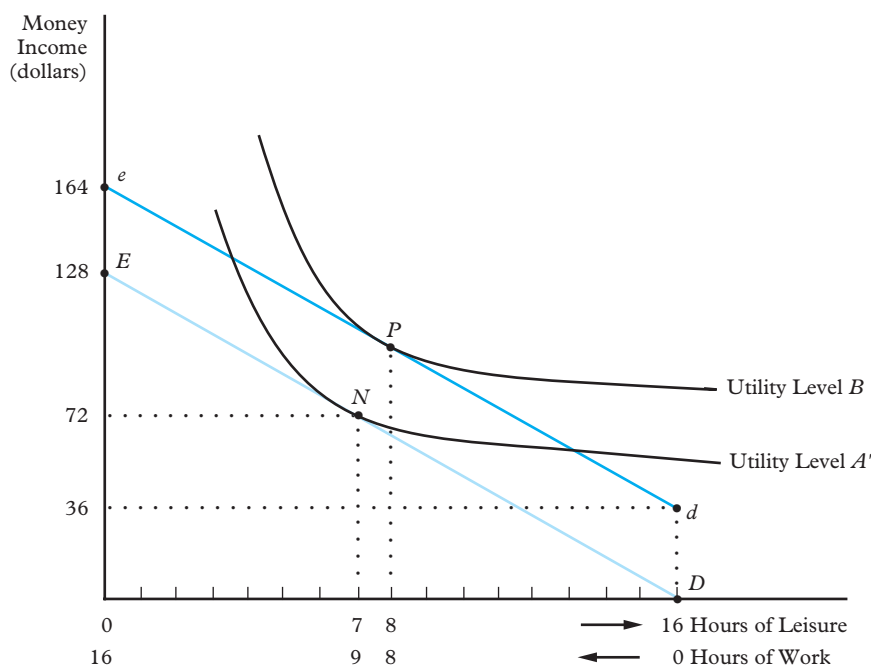
Workers who face the same budget constraint, but who have different preferences for leisure, will make different choices about hours of work. If the person whose preferences were depicted in Figure 6.5 had placed lower values on leisure time—and therefore had indifference curves that were comparatively flatter, such as the one shown in Figure 6.4b—then the point of tangency with constraint ED would have been to the left of point N (indicating more hours of work). Conversely, if he or she had steeper indifference curves, signifying that leisure time was more valuable (see Figure 6.4a), then the point of tangency in Figure 6.5 would have been to the right of point N , and fewer hours of work would have been desired. Indeed, some people will have indifference curves so steep (that is, preferences for leisure so strong) that there is no point of tangency with ED . For these people, as is illustrated by Figure 6.6, utility is maximized at the “corner” (point D); they desire no work at all and therefore are not in the labor force.

The Income Effect Suppose now that the person depicted in Figure 6.5 receives a source of income independent of work. Suppose further that this *nonlabor* income amounts to about \$36 per day. Thus, even if this person worked 0 hours per day, his or her daily income would be \$36. Naturally, if the person worked more than 0 hours, his or her daily income would be equal to \$36 plus earnings (the wage multiplied by the hours of work).

Our person’s command over resources has clearly increased, as can be shown by drawing a new budget constraint to reflect the nonlabor income. As shown by the darker blue line in Figure 6.7, the endpoints of the new constraint are point d (0 hours of work and \$36 of money income) and point e (16 hours of

Figure 6.7

Indifference Curves and Budget Constraint (with an Increase in Nonlabor Income)



work and \$164 of income—\$36 in nonlabor income plus \$128 in earnings). Note that the new constraint is *parallel* to the old one. Parallel lines have the same slope; since the slope of each constraint reflects the wage rate, we can infer that the increase in nonlabor income has not changed the person's wage rate.

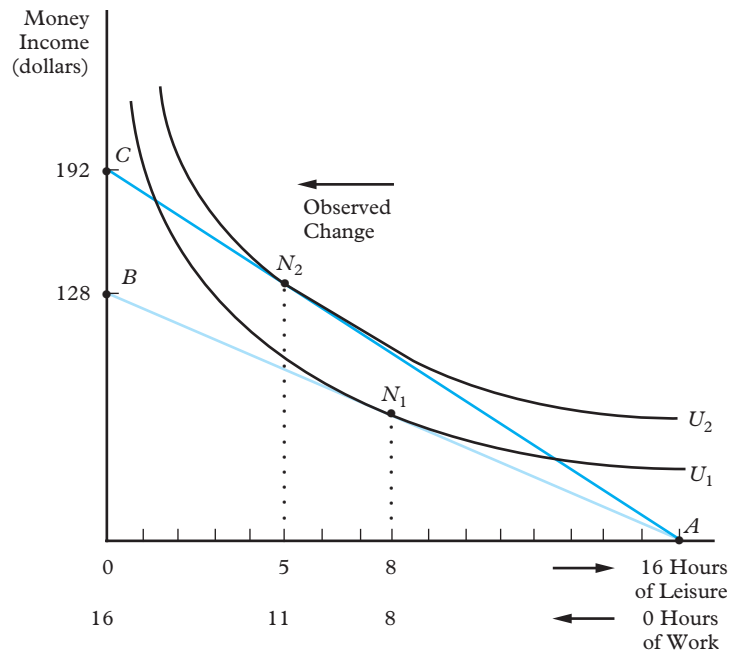
We have just described a situation in which a pure *income effect* should be observed. Income (wealth) has been increased, but the wage rate has remained unchanged. The previous section noted that if wealth *increased* and the opportunity cost of leisure remained constant, the person would consume more leisure and work *less*. We thus concluded that the income effect was negative, and this negative relationship is illustrated graphically in Figure 6.7.

When the old budget constraint (ED) was in effect, the person's highest level of utility was reached at point N , working 9 hours a day. With the new constraint (ed), the optimum hours of work are 8 per day (point P). The new source of income, because it does not alter the wage, has caused an income effect that results in one less hour of work per day. Statistical analyses of people who received large inheritances (Example 6.3) or who won large lottery prizes¹²

¹²Guido W. Imbens, Donald B. Rubin, and Bruce I. Sacerdote, "Estimating the Effects of Unearned Income on Labor Earnings, Savings, and Consumption: Evidence from a Survey of Lottery Players," *American Economic Review* 91 (September 2001): 778–794.

Figure 6.8

Wage Increase with Substitution
Effect Dominating



support the prediction that labor supply is reduced when unearned income rises.

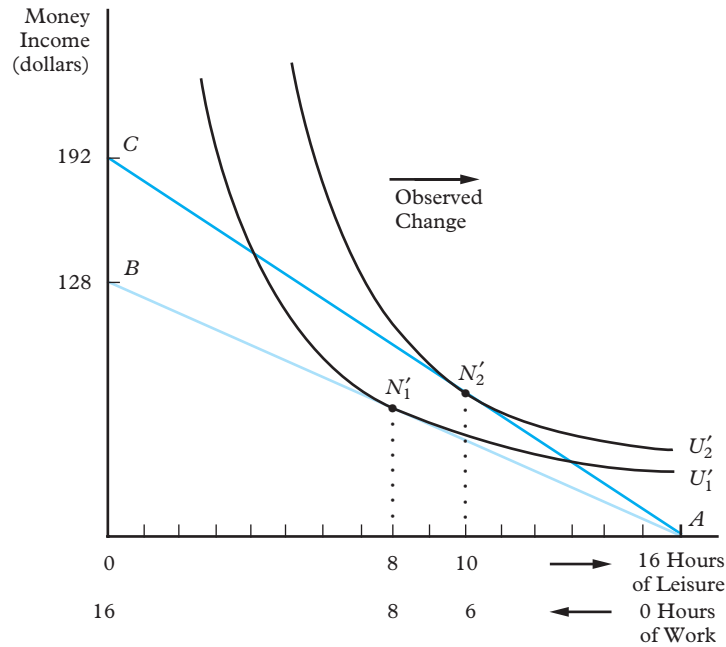
Income and Substitution Effects with a Wage Increase Suppose that instead of increasing one's command over resources by receiving a source of nonlabor income, the wage rate were to be increased from \$8 to \$12 per hour. This increase, as noted earlier, would cause *both* an income effect and a substitution effect; workers would be wealthier *and* face a higher opportunity cost of leisure. Theory tells us in this case that the substitution effect pushes them toward more hours of work and the income effect toward fewer, but it cannot tell us which effect will dominate.

Figures 6.8 and 6.9 illustrate the possible effects of the above wage change on a person's labor supply, which we now assume is initially 8 hours per day. Figure 6.8 illustrates the case in which the observed response by a worker is to increase the hours of work; in this case, the substitution effect is stronger than the income effect. Figure 6.9 illustrates the case in which the income effect is stronger and the response to a wage increase is to reduce the hours of work. The difference between the two figures lies *solely* in the shape of the indifference curves that might describe a person's preferences; the budget constraints, which reflect wealth and the wage rate, are exactly the same.

Figures 6.8 and 6.9 both show the old constraint, AB , the slope of which reflects the wage of \$8 per hour. They also show the new one, AC , which reflects

Figure 6.9

Wage Increase with
Income Effect Dominating



the \$12 wage. Because we assume workers have no source of nonlabor income, both constraints are anchored at point A , where income is zero if a person does not work. Point C on the new constraint is now at \$192 (16 hours of work times \$12 per hour).

With the worker whose preferences are depicted in Figure 6.8, the wage increase makes utility level U_2 the highest that can be reached. The tangency point at N_2 suggests that 11 hours of work is optimum. When the old constraint was in effect, the utility-maximizing hours of work were 8 per day (point N_1). Thus, the wage increase would cause this person's desired hours of work to increase by 3 per day.

With the worker whose preferences are depicted in Figure 6.9, the wage increase would make utility level U_2 the highest one possible (the prime emphasizes that workers' preferences differ and that utility *levels* in Figures 6.8 and 6.9 cannot be compared). Utility is maximized at N'_2 , at 6 hours of work per day. Thus, with preferences like those in Figure 6.9, working hours fall from 8 to 6 as the wage rate increases.

Isolating Income and Substitution Effects We have graphically depicted the income effect by itself (Figure 6.7) and the two possible outcomes of an increase in wages (Figures 6.8 and 6.9), which combine the income and substitution effects. Is it possible to graphically isolate the substitution effect? The answer is yes, and

EXAMPLE 6.3**Do Large Inheritances Induce Labor Force Withdrawal?**

Do large bequests of unearned income reduce people's incentives to work? One study divided people who received inheritances in 1982–1983 into two groups: those who received small bequests (averaging \$7,700) and those who received larger ones, averaging \$346,200. The study then analyzed changes in the labor force participation behavior of the two groups between 1982 and 1985. Not surprisingly, those who received the larger inheritances were more likely to drop out of the labor force. Specifically, in an environment in which other forces were causing the labor force participation rate among the small-bequest group to rise

from 76 percent to 81 percent, the rate in the large-bequest group fell from 70 percent to 65 percent. Somewhat more surprising was the fact that perhaps in anticipation of the large bequest, the labor force participation rate among the people in the latter group was lower to begin with!

Data from: Douglas Holtz-Eakin, David Joulfaian, and Harvey S. Rosen, "The Carnegie Conjecture: Some Empirical Evidence," *Quarterly Journal of Economics* 108, no. 2 (1993): 413–435. The findings reported above hold up even after controlling for such factors as age and earnings.

the most meaningful way to do this is to return to the context of a wage change, such as the one depicted in Figures 6.8 and 6.9. We arbitrarily choose to analyze the response shown in Figure 6.8.

Figure 6.10 has three panels. Panel (a) repeats Figure 6.8; it shows the final, overall effect of a wage increase on the labor supply of the person whose preferences are depicted. As we saw earlier, the effect of the wage increase in this case is to raise the person's utility from U_1 to U_2 and to induce this worker to increase desired hours of work from 8 to 11 per day. Embedded in this overall effect of the wage increase, however, is an income effect pushing toward less work and a substitution effect pushing toward more. These effects are graphically separated in panels (b) and (c).

Panel (b) of Figure 6.10 shows the income effect that is embedded in the overall response to the wage change. By definition, the income effect is the change in desired hours of work brought on by increased wealth, holding the wage rate constant. To reveal this embedded effect, we ask a hypothetical question: "What would have been the change in labor supply if the person depicted in panel (a) had reached the new indifference curve (U_2) with a change in *nonlabor* income instead of a change in his or her wage rate?"

We begin to answer this question graphically by moving the old constraint to the northeast, which depicts the greater command over leisure time and goods—and hence the higher level of utility—associated with greater wealth. The constraint is shifted outward while maintaining its original slope (reflecting the old \$8 wage), which holds the wage constant. The dashed line in panel (b), which is parallel to AB , depicts this hypothetical movement of the old constraint, and it results in a tangency point at N_3 . This tangency suggests that had the person received nonlabor income, with no change in the wage, sufficient to reach the new level of utility, he or she would have *reduced* work hours from 8 (N_1) to 7 (N_3) per

day. This shift is graphical verification that the income effect is negative, assuming that leisure is a normal good.

The substitution effect is the effect on labor supply of a change in the wage rate, holding wealth constant. It can be seen in panel (c) of Figure 6.10 as the difference between where the person actually ended up on indifference curve U_2 (tangency at N_2) and where he or she would have ended up with a pure income effect (tangency at N_3). Comparing tangency points on the *same* indifference curve is a graphical approximation to holding wealth constant. Thus, *with* the wage change, the person represented in Figure 6.10 ended up at point N_2 , working 11 hours a day. *Without* the wage change, the person would have chosen to work 7 hours a day (point N_3). The wage change *by itself*, holding utility (or real wealth) constant, caused work hours to increase by 4 per day.¹³ This increase demonstrates that the substitution effect is positive.

To summarize, the observed effect of raising wages from \$8 to \$12 per hour increased the hours of work in Figure 6.10 from 8 to 11 per day. This observed effect, however, is the *sum* of two component effects. The income effect, which operates because a higher wage increases one's real wealth, tended to *reduce* the hours of work from 8 to 7 per day. The substitution effect, which captures the pure effect of the change in leisure's opportunity cost, tended to push the person toward 4 more hours of work per day. The end result was an increase of 3 in the hours worked each day.

Which Effect Is Stronger? Suppose that a wage increase changes the budget constraint facing a worker from CD to CE in Figure 6.11. If the worker had a relatively flat set of indifference curves, the initial tangency along CD might be at point A , implying a relatively heavy work schedule. If the person had more steeply sloped indifference curves, the initial tangency might be at point B , where hours at work are fewer.

One important influence on the size of the income effect is the extent of the northeast movement of the new constraint: the more the constraint shifts outward, the greater the income effect will tend to be. For a person with an initial tangency at point A , for example, the northeast movement is larger than that for a person whose initial tangency is at point B . Put in words, the increased command over resources made possible by a wage increase is only attainable if one works, and the more work-oriented the person is, the greater will be his or her increase in resources. Other things equal, people who are working longer hours will exhibit greater income effects when wage rates change.

To take this reasoning to the extreme, suppose a person's indifference curves were so steep that the person was initially out of the labor force (that is, when the

¹³In our initial definition of the substitution effect, we held *money income* constant, while in the graphical analysis, we held *utility* constant. These slightly different approaches were followed for explanatory convenience, and they represent (respectively) the theoretical analyses suggested by Evgeny Slutsky and John Hicks. For an easy-to-follow explanation of the two approaches, see Heinz Kohler, *Intermediate Microeconomics* (Glenview, Ill.: Scott Foresman, 1986): 76–81.

Figure 6.10

Wage Increase with Substitution Effect
Dominating: Isolating Income and Substitution
Effects

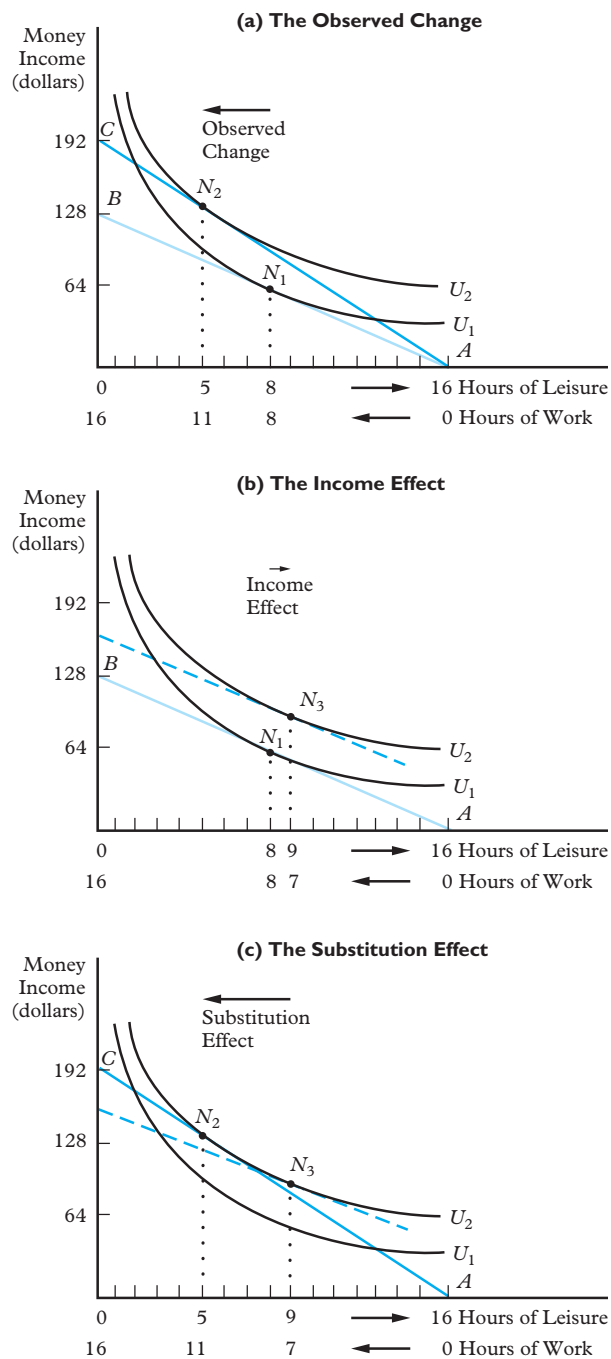
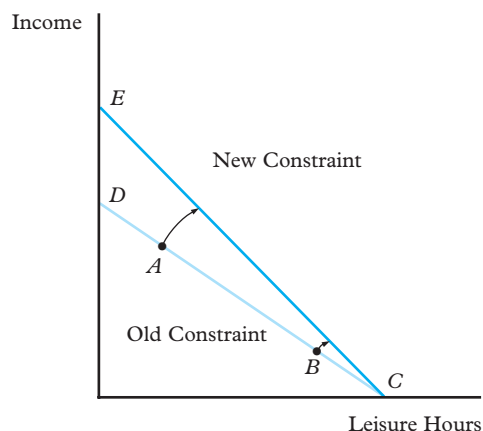


Figure 6.11

The Size of the Income Effect Is Affected by the Initial Hours of Work



budget constraint was CD in Figure 6.11, his or her utility was maximized at point C . The wage increase and the resultant new constraint, CE , can induce only two outcomes: the person will either begin to work for pay or remain out of the labor force. *Reducing* the hours of paid employment is not possible. For those who are out of the labor force, then, the decision to *participate* as wage offers rise clearly reflects a dominant substitution effect. Conversely, if someone currently working decides to change his or her participation decision and drop out of the labor force when wages fall, the substitution effect has again dominated. Thus, the *labor force participation decisions brought about by wage changes exhibit a dominant substitution effect*. We turn now to a more detailed analysis of the decision whether to join the labor force.

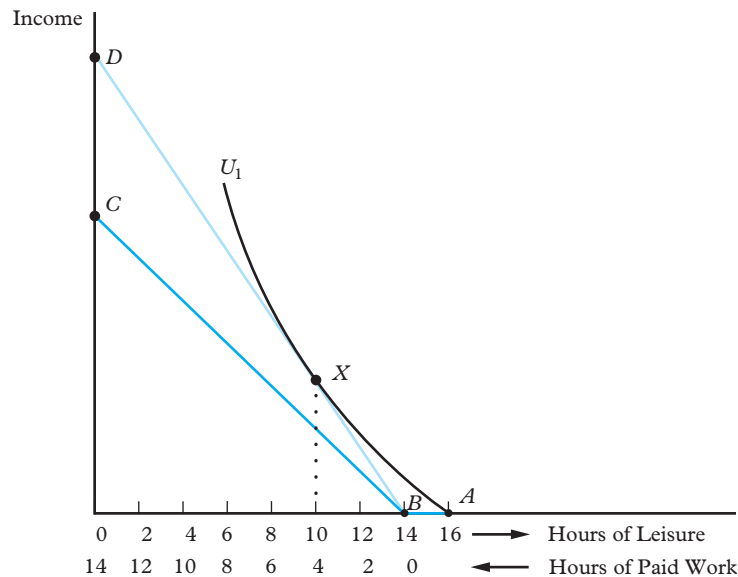
The Reservation Wage An implication of our labor supply theory is that if people who are not in the labor force place a value of $\$X$ on the marginal hour of leisure, then they would be unwilling to take a job unless the offered wages were greater than $\$X$. Because they will “reserve” their labor unless the wage is $\$X$ or more (see Example 6.4), economists say that they have a *reservation wage* of $\$X$. The reservation wage, then, is the wage below which a person will not work, and in the labor/leisure context, it represents the value placed on an hour of lost leisure time.¹⁴

Refer back to Figure 6.6, which graphically depicted a person choosing not to work. The reason there was no tangency between an indifference curve and

¹⁴See Hans G. Bloemen and Elena G. F. Stancanelli, “Individual Wealth, Reservation Wages, and Transitions into Employment,” *Journal of Labor Economics* 19 (April 2001): 400–439, for a study of reservation wages.

Figure 6.12

Reservation Wage with Fixed Time
Costs of Working



the budget constraint—and the reason the person remained out of the labor force—was that the wage was everywhere lower than his or her marginal value of leisure time.

Often, people are thought to behave as if they have both a reservation wage and a certain number of work hours that must be offered before they will consider taking a job. The reasons are not difficult to understand and are illustrated in Figure 6.12. Suppose that taking a job entails 2 hours of commuting time (round-trip) per day. These hours, of course, are unpaid, so the worker's budget constraint must reflect that if a job is accepted, 2 hours of leisure are given up before there is any increase in income. These fixed costs of working are reflected in Figure 6.12 by segment AB . Segment BC , of course, reflects the earnings that are possible (once at work), and the slope of BC represents the person's wage rate.

Is the wage underlying BC great enough to induce the person to work? Consider indifference curve U_1 , which represents the highest level of utility this person can achieve, given budget constraint ABC . Utility is maximized at point A , and the person chooses not to work. It is clear from this choice that the offered wage (given the 2-hour commute) is below the person's reservation wage, but can we show the latter wage graphically?

To take work with a 2-hour commute, the person depicted in Figure 6.12 must find a job able to generate a combination of earnings and leisure time that yields a utility level equal to, or greater than, U_1 . This is possible only if the person's budget constraint is equal to (or to the right of) ABD , which is tangent to U_1

EXAMPLE 6.4**Daily Labor Supply at the Ballpark**

The theory of labor supply rests in part on the assumption that when workers' offered wages climb above their reservation wages, they will decide to participate in the labor market. An implication of this theory is that in jobs for which hiring is done on a daily basis, and for which wages fluctuate widely from day to day, we should observe daily fluctuations in participation. These expectations are supported by the daily labor supply decisions of vendors at Major League Baseball games.

One such study examined the individual labor supply behavior of vendors in one ballpark over the course of the 1996 major league baseball season. Vendors walk through the stands selling food and drinks, and their earnings are completely determined by the sales they are able to make each day. The vendors studied could freely choose whether to work any given game, and the data collected by this study clearly suggest they made their decisions by weighing their opportunity cost of working against their expected earnings during the course of the

game. (Expected earnings, of course, are related to a number of factors, including how many fans were likely to attend the game.)

The study was able to compare the actual amount earned by each vendor at each game with the number of vendors who had decided to work. The average amount earned by vendors was \$43.81, with a low of \$26.55 for one game and a high of \$73.15 for another—and about 45 vendors worked the typical game at this ballpark. The study found that an increase in average earnings of \$10 (which represents about a one standard deviation increase from the mean of \$43.81) lured about six extra vendors to the stadium.

Clearly, then, vendors behaved as if they had reservation wages that they compared with expected earnings when deciding whether to work particular games.

Data from: Gerald Oettinger, "An Empirical Analysis of the Daily Labor Supply of Stadium Vendors," *Journal of Political Economy* 107 (April 1999): 360–392.

at point X. The person's reservation wage, then, is equal to the slope of BD , and you can readily note that in this case, the slope of BD exceeds the slope of BC , which represents the currently offered wage. Moreover, to bring utility up to the level of U_1 (the utility associated with not working), the person shown in Figure 6.12 must be able to find a job at the reservation wage that offers 4 hours of work per day. Put differently, at this person's reservation wage, he or she would want to consume 10 hours of leisure daily, and with a 2-hour commute, this implies 4 hours of work.

Empirical Findings on the Income and Substitution Effects

Labor supply theory suggests that the choices workers make concerning their desired hours of work depend on their wealth and the wage rate they can command, in addition to their preferences. In particular, this theory suggests the existence of a negative income effect and a positive substitution effect. Empirical tests of labor supply theory generally attempt to determine if these two effects can be observed, if they operate in the expected directions, and what their relative magnitudes are.

Most recent studies of labor supply have used large samples of individuals to analyze how labor force participation and hours of work are affected by wage rates and income, holding other influences (age, for example) constant. Studies of male and female labor force behavior are done separately because of the different roles men and women typically play in performing household work and child-rearing—activities that clearly affect labor supply decisions but about which information is usually very limited.

The studies of labor supply behavior for men between the ages of 25 and 55 generally conclude that both income and substitution effects are small, perhaps even zero. Probably because the *net* responses to wage changes are so close to zero, the results of studies that try to separately measure the income and substitution effects—while generally supportive of the theory—are highly dependent on the statistical methods used.¹⁵ Studies of older men tend to focus on retirement behavior (a topic we will address in chapter 7) and find, as theory suggests, that the substitution effect dominates the decision whether to withdraw from the labor force. In particular, the sharp rise in early retirements in the last two decades of the twentieth century was concentrated among men with lower levels of education, for whom wages fell during that period.¹⁶

Studies of the labor supply behavior of married women generally have found a greater responsiveness to wage changes than is found among men, and recent work suggests two generalizations. First, changes in the *hours of work* associated with a wage change for married women are closer to those for men than are changes in *labor force participation*; that is, as seen in Example 6.5, the labor force participation rate for married women has been more responsive to wage changes than have been the hours of work. Second, in the last two decades, the labor supply behavior of married women has become much more similar to that for men—meaning that the labor supply of women is *becoming less responsive to wage changes* than it used to be. The reduced responsiveness has been especially noticeable in women's labor force participation decisions, where the differences between men and women have been greatest.¹⁷ This growing similarity in labor supply behavior may well reflect a growing similarity in the expectations held by women and men concerning work and careers.

¹⁵Matias Eklof and Hans Sacklen, "The Hausman-MaCurdy Controversy: Why Do the Results Differ Across Studies," *Journal of Human Resources* 35 (Winter 2000): 204–220; and James P. Ziliak and Thomas J. Kniesner, "The Effect of Income Taxation on Consumption and Labor Supply," *Journal of Labor Economics* 23 (October 2005): 769–796.

¹⁶Franco Peracchi and Finis Welch, "Trends in the Labor Force Transitions of Older Men and Women," *Journal of Labor Economics* 12 (April 1994): 210–242.

¹⁷Francine D. Blau and Lawrence M. Kahn, "Changes in the Labor Supply Behavior of Married Women: 1980–2000," *Journal of Labor Economics* 25 (July 2007): 393–438; Bradley T. Heim, "Structural Estimation of Family Labor Supply with Taxes: Estimating a Continuous Hours Model Using a Direct Utility Specification," *Journal of Human Resources* 44 (Spring 2009): 350–385; Kelly Bishop, Bradley Heim, and Kata Mihaly, "Single Women's Labor Supply Elasticities: Trends and Policy Implications," *Industrial and Labor Relations Review* 63 (October 2009): 146–168.

EXAMPLE 6.5**Labor Supply Effects of Income Tax Cuts**

In 1986, Congress changed the personal income tax system in the United States by drastically reducing tax rates on upper levels of income. Before this change, for example, families paid a 50 percent tax rate on taxable incomes over \$170,000; after the change, this tax rate was reduced to 28 percent. The tax rate on taxable incomes over \$50,000 was also set at 28 percent, down from about 40 percent. Lower income tax rates have the effect of increasing take-home earnings, and they therefore act as an increase in wage rates. Because lower rates generate an income and a substitution effect that work in opposite directions, they have an ambiguous anticipated effect on labor supply. Can we find out which effect is stronger in practice?

The 1986 changes served as a *natural experiment* (abrupt changes in only one variable, the sizes of which vary by group). The changes were sudden, large, and very different for families of different incomes. For married women in families that, without their earnings, had incomes at the 99th percentile of the income distribution (that is, the upper 1 percent), the tax rate cuts meant a 22 percent increase in their take-home wage rates. For women in families with incomes at the 90th percentile, the smaller tax rate cuts meant a 12 percent increase in take-home wages. It turns out that married women at the 99th and 90th percentiles of family income were similar in age, education, and occupation—and increases in their labor supply had

been similar prior to 1986. Therefore, comparing their responses to very different changes in their after-tax wage rates should yield insight into how the labor supply of married women responded to tax rate changes.

One study compared labor supply increases, from 1984 to 1990, for married women in the 99th and 90th percentiles. It found that the labor force participation rate for women in the 99th percentile rose by 19.4 percent and that, if working, their hours of work rose by 12.7 percent during that period. In contrast, both labor force participation and hours of work for women at the 90th percentile rose only by about 6.5 percent. The data from this natural experiment, then, suggest that women who experienced larger increases in their take-home wages desired greater increases in their labor supply—which implies that the substitution effect dominated the income effect for these women. Also, consistent with both theory and the results from other studies (discussed in the text), the dominance of the substitution effect was more pronounced for labor force participation decisions than it was for hours-of-work decisions.

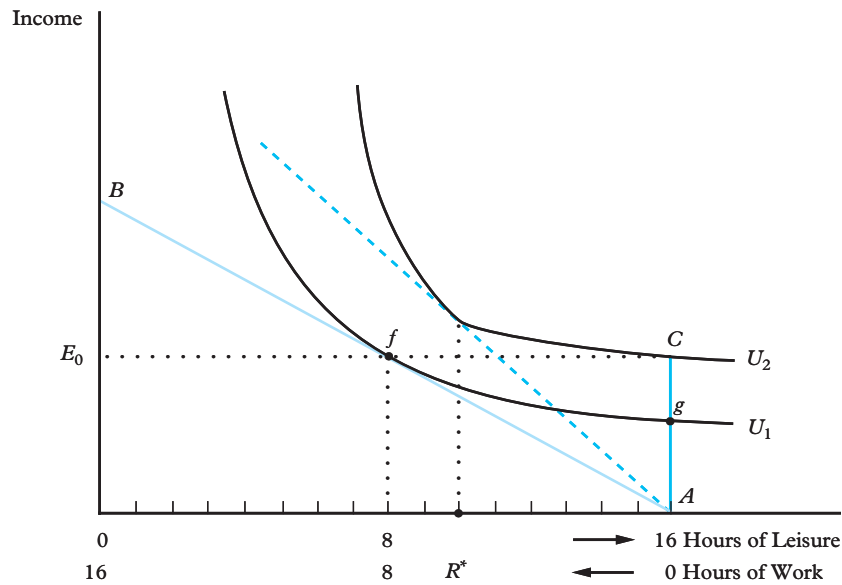
Data from: Nada Eissa, "Taxation and Labor Supply of Married Women: The Tax Reform Act of 1986 as a Natural Experiment," Working Paper No. 5023, National Bureau of Economic Research, Cambridge, Mass., February 1995.

Policy Applications

Many income maintenance programs create budget constraints that increase income while reducing the take-home wage rate (thus causing the income and substitution effects to work in the same direction). Therefore, using labor supply theory to analyze the work-incentive effects of various social programs is both instructive and important. We characterize these programs by the budget constraints they create for their recipients.

Figure 6.13

Budget Constraint with a Spike



Budget Constraints with “Spikes”

Some social insurance programs compensate workers who are unable to work because of a temporary work injury, a permanent disability, or a layoff. Workers' compensation insurance replaces most of the earnings lost when workers are hurt on the job, and private or public disability programs do the same for workers who become physically or emotionally unable to work for other reasons. Unemployment compensation is paid to those who have lost a job and have not been able to find another. While exceptions can be found in the occasional jurisdiction,¹⁸ it is generally true that these *income replacement* programs share a common characteristic: they pay benefits only to those who are not working.

To understand the consequences of paying benefits only to those who are not working, let us suppose that a workers' compensation program is structured so that, after injury, workers receive their pre-injury earnings for as long as they are off work. Once they work even one hour, however, they are no longer considered disabled and cannot receive further benefits. The effects of this program on work incentives are analyzed in Figure 6.13, in which it is assumed that the pre-injury budget constraint was AB and pre-injury earnings were $E_0 (= AC)$.

¹⁸UI and workers' compensation programs in the United States are run at the state level and thus vary in their characteristics to some extent.

Furthermore, we assume that the worker's "market" budget constraint (that is, the constraint in the absence of a workers' compensation program) is unchanged, so that after recovery, the pre-injury wage can again be earned. Under these conditions, the post-injury budget constraint is BAC , and the person maximizes utility at point C —a point of no work.

Note that constraint BAC contains the segment AC , which looks like a spike. It is this spike that creates severe work-incentive problems, for two reasons. First, the returns associated with the first hour of work are *negative*. That is, a person at point C who returns to work for 1 hour would find his or her income to be considerably reduced by working. Earnings from this hour of work would be more than offset by the reduction in benefits, which creates a negative "net wage." The substitution effect associated with this program characteristic clearly discourages work.¹⁹

Second, our assumed no-work benefit of AC is equal to E_0 , the pre-injury level of earnings. If the worker values leisure at all (as is assumed by the standard downward slope of indifference curves), being able to receive the old level of earnings while also enjoying more leisure clearly enhances utility. The worker is better off at point C than at point f , the pre-injury combination of earnings and leisure hours, because he or she is on indifference curve U_2 rather than U_1 . Allowing workers to reach a higher utility level without working generates an income effect that discourages, or at least slows, the return to work.

Indeed, the program we have assumed raises a worker's reservation wage above his or her pre-injury wage, meaning that a return to work is possible only if the worker qualifies for a higher-paying job. To see this graphically, observe the dashed blue line in Figure 6.13 that begins at point A and is tangent to indifference curve U_2 (the level of utility made possible by the social insurance program). The slope of this line is equal to the person's reservation wage, because if the person can obtain the desired hours of work at this or a greater wage, utility will be at least equal to that associated with point C . Note also that for labor force participation to be induced, the reservation wage must be received for at least R^* hours of work.

Given that the work-incentive aspects of income replacement programs often quite justifiably take a backseat to the goal of making unfortunate workers "whole" in some economic sense, creating programs that avoid work disincentives is not easy. With the preferences of the worker depicted in Figure 6.13, a benefit of slightly less than A_g would ensure minimal loss of utility while still

¹⁹In graphical terms, the budget constraint contains a vertical spike, and the slope of this vertical segment is infinitely negative. In economic terms, the implied infinitely negative (net) wage arises from the fact that even 1 minute of work causes a person to lose his or her entire benefit. For empirical evidence, see Susan Chen and Wilbert van der Klaauw, "The Work Disincentive Effects of the Disability Insurance Program in the 1990s," *Journal of Econometrics* 142 (February 2008): 757–784. For an analysis of disability insurance usage, see David H. Autor and Mark G. Duggan, "The Growth in the Social Security Disability Rolls: A Fiscal Crisis Unfolding," *Journal of Economic Perspectives* 20 (Summer 2006): 71–96.

EXAMPLE 6.6**Staying Around One's Kentucky Home: Workers' Compensation Benefits and the Return to Work**

Workers injured on the job receive workers' compensation insurance benefits while away from work. These benefits differ across states, but they are calculated for most workers as some fraction (normally two-thirds) of weekly, pre-tax earnings. For high-wage workers, however, weekly benefits are typically capped at a maximum, which again varies by state.

On July 15, 1980, Kentucky raised its maximum weekly benefit by 66 percent. It did not alter benefits in any other way, so this change effectively granted large benefit increases to high-wage workers without awarding them to anyone else. Because those injured before July 15 were ineligible for the increased benefits, even if they remained off work after July 15, this policy change created a nice natural experiment: one group of injured workers was

able to obtain higher benefits, while another group was not. Did the group receiving higher benefits show evidence of reduced labor supply, as suggested by theory?

The effects of increased benefits on labor supply were unmistakable. High-wage workers ineligible for the new benefits typically stayed off the job for four weeks, but those injured after July 15 stayed away for five weeks—25 percent longer! No increases in the typical time away from work were recorded among lower-paid injured workers, who were unaffected by the changes in benefits.

Data from: Bruce D. Meyer, W. Kip Viscusi, and David L. Durbin, "Workers' Compensation and Injury Duration: Evidence from a Natural Experiment," *American Economic Review* 85 (June 1995): 322–340.

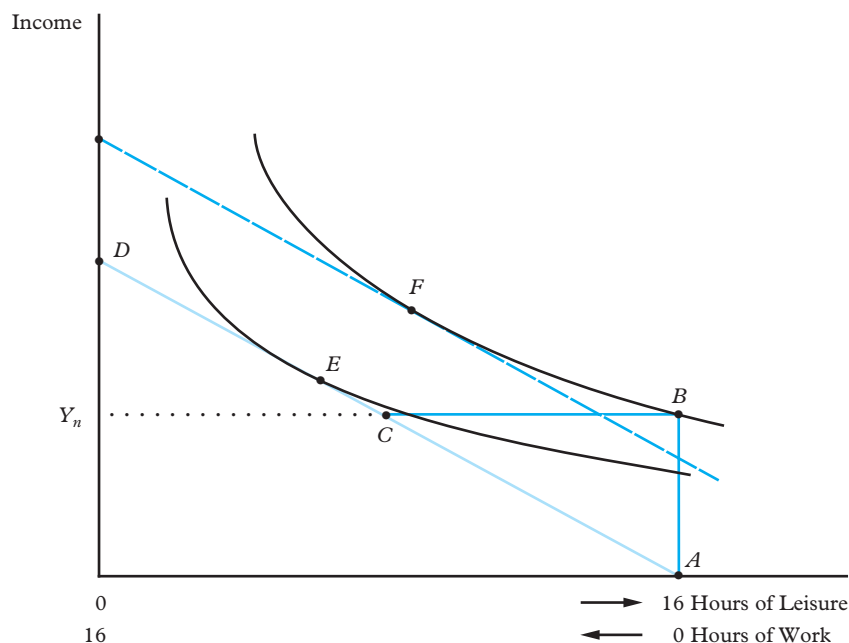
providing incentives to return to work as soon as physically possible (work would allow indifference curve U_1 to be attained—see point f —while not working and receiving a benefit of less than Ag would not). Unfortunately, workers differ in their preferences, so the optimal benefit—one that would provide work incentives yet ensure only minimal loss of utility—differs for each individual.

With programs that create spikes, the best policymakers can do is set a no-work benefit as some fraction of previous earnings and then use administrative means to encourage the return to work among any whose utility is greater when not working. Unemployment insurance (UI), for example, replaces something like half of lost earnings for the typical worker, but the program puts an upper limit on the weeks each unemployed worker can receive benefits. Workers' compensation replaces two-thirds of lost earnings for the average worker but must rely on doctors—and sometimes judicial hearings—to determine whether a worker continues to be eligible for benefits. (For evidence that more-generous workers' compensation benefits do indeed induce longer absences from work, see Example 6.6.)²⁰

²⁰For a summary of evidence on the labor supply effects of UI and workers' compensation, see Alan B. Krueger and Bruce D. Meyer, "Labor Supply Effects of Social Insurance," in *Handbook of Public Economics*, vol. 4, eds. Alan Auerbach and Martin Feldstein (Amsterdam: North Holland, 2002); and Peter Kuhn and Chris Riddell, "The Long-Term Effects of Unemployment Insurance: Evidence from New Brunswick and Maine, 1940–1991," *Industrial and Labor Relations Review* 63 (January 2010): 183–204.

Figure 6.14

Income and Substitution
Effects for the Basic
Welfare System



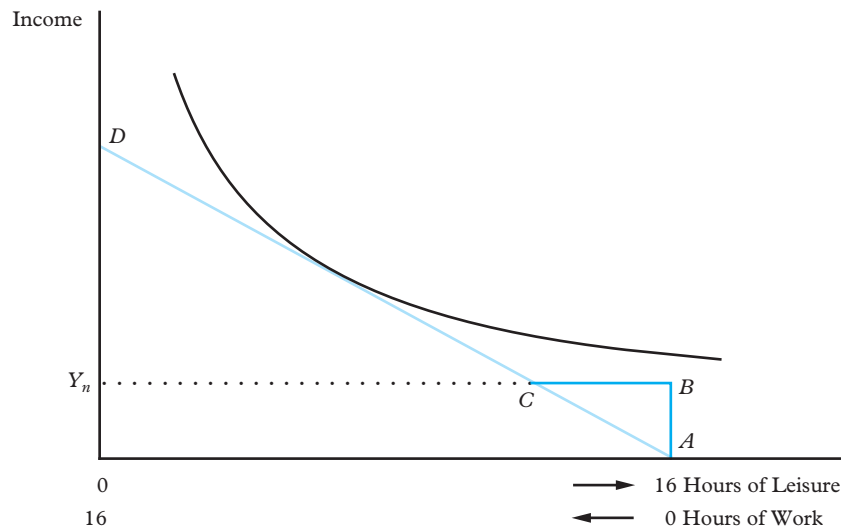
Programs with Net Wage Rates of Zero

The programs just discussed were intended to confer benefits on those who are unable to work, and the budget-constraint spike was created by the eligibility requirement that to receive benefits, one must not be working. Other social programs, such as welfare, have different eligibility criteria and calculate benefits differently. These programs factor income needs into their eligibility criteria and then pay benefits based on the difference between one's actual earnings and one's needs. We will see that paying people the difference between their earnings and their needs creates a net wage rate of zero; thus, the work-incentive problems associated with these welfare programs result from the fact that they increase the income of program recipients while also drastically reducing the price of leisure.

Nature of Welfare Subsidies Welfare programs have historically taken the form of a guaranteed annual income, under which the welfare agency determines the income needed by an eligible person (Y_n in Figure 6.14) based on family size, area living costs, and local welfare regulations. Actual earnings are then subtracted from this needed level, and a check is issued to the person each month for the difference. If the person does not work, he or she receives a subsidy of Y_n . If the person works, and if earnings cause dollar-for-dollar reductions in welfare benefits, then a budget constraint like $ABCD$ in Figure 6.14 is created. The person's income

Figure 6.15

The Basic Welfare System:
A Person Not Choosing
Welfare



remains Y_n as long as he or she is subsidized. If receiving the subsidy, then, an extra hour of work yields *no* net increase in income, because the extra earnings result in an equal reduction in welfare benefits. The net wage of a person on the program—and therefore his or her price of leisure—is zero, which is graphically shown by the segment of the constraint having a slope of zero (BC).²¹

Thus, a welfare program like the one summarized in Figure 6.14 increases the income of the poor by moving the lower end of the budget constraint out from AC to ABC ; as indicated by the dashed hypothetical constraint in Figure 6.14, this shift creates an *income effect* tending to reduce labor supply from the hours associated with point E to those associated with point F . However, it *also* causes the wage to effectively drop to zero; every dollar earned is matched by a dollar reduction in welfare benefits. This dollar-for-dollar reduction in benefits induces a huge *substitution effect*, causing those accepting welfare to reduce their hours of work to zero (point B). Of course, if a person's indifference curves were sufficiently flat so that the curve tangent to segment CD passed *above* point B (see Figure 6.15), then that person's utility would be maximized by choosing work instead of welfare.²²

²¹Gary Burtless, "The Economist's Lament: Public Assistance in America," *Journal of Economic Perspectives* 4 (Winter 1990): 57–78, summarizes a variety of public assistance programs in the United States prior to 1990. This article suggests that in actual practice, benefits were usually reduced by something less than dollar for dollar (perhaps by 80 or 90 cents per dollar of earnings).

²²See Robert Moffitt, "Incentive Effects of the U.S. Welfare System: A Review," *Journal of Economic Literature* 30 (March 1992): 1–61, for a summary of the literature on labor supply effects of the welfare system.

Welfare Reform In light of the disincentives for work built into traditional welfare programs, the United States adopted major changes to its come-subsidy programs in the 1990s. The Personal Responsibility and Work Opportunity Reconciliation Act (PRWORA) of 1996 gave states more authority over how they could design their own welfare programs, with the intent of leading to more experimentation in program characteristics aimed at encouraging work, reducing poverty, and moving people off welfare.²³ PRWORA also placed a five-year (lifetime) time limit on the receipt of welfare benefits and required that after two years on welfare, recipients must work at least 30 hours per week. These changes appear to have had the effect of increasing the labor force participation rates of single mothers (the primary beneficiaries of the old welfare system); the participation rate for single mothers jumped from 68 percent in 1994 to roughly 78 percent in 2000—a much larger increase than was observed for other groups of women.²⁴

Lifetime Limits Both lifetime limits and work requirements can be analyzed using the graphical tools developed in this chapter. Lifetime limits on the receipt of welfare have the effect of ending eligibility for transfer payments, either by forcing recipients off welfare or by inducing them to leave so they can “save” their eligibility in case they need welfare later in life. Thus, in terms of Figure 6.14, the lifetime limit ultimately removes ABC from the potential recipient’s budget constraint, which then reverts to the market constraint of AD .

Clearly, the lifetime limit increases work incentives by ultimately eliminating the income subsidy. However, within the limits of their eligible years, potential welfare recipients must choose when to receive the subsidy and when to “save” their eligibility in the event of a future need. Federal law provides for welfare subsidies only to families with children under the age of 18; consequently, the closer one’s youngest child is to 18 (when welfare eligibility ends anyway), the smaller are the incentives of the parent to forgo the welfare subsidy and save eligibility for the future.²⁵

Work Requirements As noted earlier, PRWORA introduced a work requirement into the welfare system, although in some cases, unpaid work or enrolling in education or training programs counts toward that requirement. States differ in how the earnings affect welfare benefits, and many have rules that allow welfare recipients to keep most of what they earn (by not reducing, at least by much, their welfare benefits); we analyze such programs in the next section. For now, we can understand the *basic* effects of a work requirement by maintaining our assumption that earnings reduce welfare benefits dollar for dollar.

²³For a comprehensive summary of the reforms and various analyses of them, see Jeffrey Grogger and Lynn A. Karoly, *Welfare Reform: Effects of a Decade of Change* (Cambridge, Mass.: Harvard University Press, 2005).

²⁴Rebecca M. Blank, “Evaluating Welfare Reform in the United States,” *Journal of Economic Literature* 40 (December 2002): 1105–1166.

²⁵Jeffrey Grogger, “Time Limits and Welfare Use,” *Journal of Human Resources* 39 (Spring 2004): 405–424.

Figure 6.16

The Welfare System with a Work Requirement

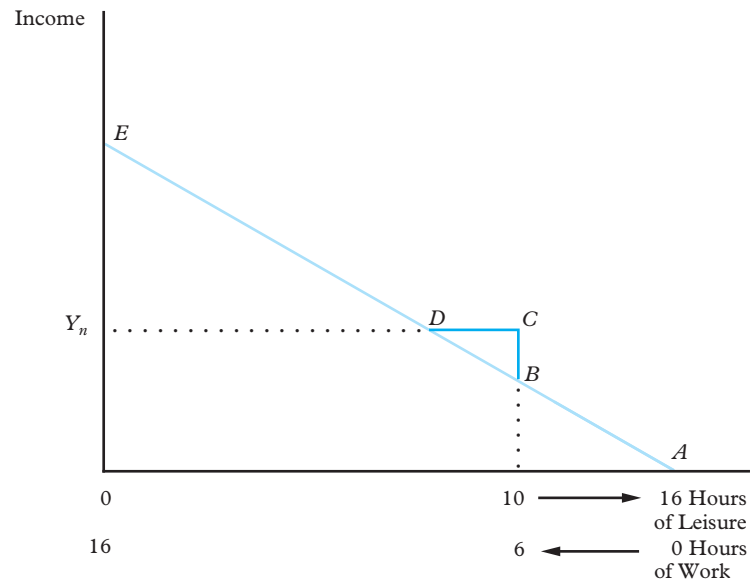


Figure 6.16 illustrates the budget constraint associated with a minimum work requirement of 6 hours a day (30 hours per week). If the person fails to work the required 6 hours a day, no welfare benefits are received, and he or she will be along segment AB of the constraint. If the work requirement is met, but earnings are less than Y_n , welfare benefits are received (see segment BCD). If the work requirement is exceeded, income (earnings plus benefits) remains at Y_n —the person is along CD —until earnings rise above needed income and the person is along segment DE of the constraint and no longer eligible for welfare benefits.

The work-incentive effects of this *work requirement* can be seen from analyzing Figure 6.16 in the context of people whose skills are such that they are potential welfare recipients. At one extreme, some potential recipients may have such steeply sloped indifference curves (reflecting a strong preference, or a need, to stay at home) that utility is maximized along segment AB , where so little market work is performed that they do not qualify for welfare. At the opposite extreme, others may have such flat indifference curves (reflecting a strong preference for income and a weak preference for leisure) that their utility is maximized along segment DE ; they work so many hours that their earnings disqualify them for welfare benefits.

In the middle of the above extremes will be those whose preferences lead them to work enough to qualify for welfare benefits. Clearly, if their earnings reduce their benefits dollar for dollar—as shown by the horizontal segment DC in Figure 6.16—they will want to work just the *minimum hours* needed to qualify for welfare, because their utility will be maximized at point C and not along DC . (For

labor supply responses to different forms of a work requirement—requisitions of food from farmers during wartime—see Example 6.7 on page 204.)

Subsidy Programs with Positive Net Wage Rates

So far, we have analyzed the work-incentive effects of income maintenance programs that create net wage rates for program recipients that are either negative or zero (that is, they create constraints that have either a spike or a horizontal segment). Most current programs, however, including those adopted by states under PRWORA, create positive net wages for recipients. Do these programs offer a solution to the problem of work incentives? We will answer this question by analyzing a relatively recent and rapidly growing program: the Earned Income Tax Credit (EITC).

The EITC program makes income tax credits available to low-income families with at least one worker. A tax credit of \$1 reduces a person's income taxes by \$1, and in the case of the EITC, if the tax credit for which workers qualify exceeds their total income tax liability, the government will mail them a check for the difference. Thus, the EITC functions as an earnings subsidy, and because the subsidy goes only to those who work, the EITC is seen by many as an income maintenance program that preserves work incentives. This view led Congress to vastly expand the EITC under President Bill Clinton and it is now the largest cash subsidy program directed at low-income households with children.

The tax credits offered by the EITC program vary with one's earnings and the number of dependent children. For purposes of our analysis, which is intended to illustrate the work-incentive effects of the EITC, we will focus on the credits in the year 2009 offered to unmarried workers with two children. Figure 6.17 graphs the relevant program characteristics for a worker with two children who could earn a market (unsubsidized) wage reflected by the slope of AC . As we will see later, for such a worker, the EITC created a budget constraint of $ABDEC$.

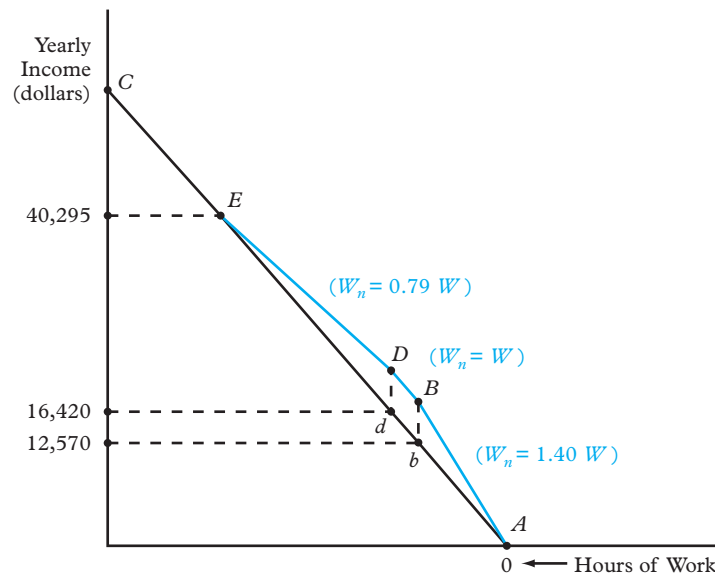
For workers with earnings of \$12,570 or less, the tax credit was calculated at 40 percent of earnings. That is, for every dollar earned, a tax credit of 40 cents was also earned; thus, for those with earnings of under \$12,570, net wages (W_n) were 40 percent higher than market wages (W). Note that this tax credit is represented by segment AB on the EITC constraint in Figure 6.17 and that the slope of AB exceeds the slope of the market constraint AC .

The maximum tax credit allowed for a single parent with two children was \$5,028 in 2009. Workers who earned between \$12,570 and \$16,420 per year qualified for this maximum tax credit. Because these workers experienced no increases or reductions in tax credits per added dollar of earnings, their net wage is equal to their market wage. The constraint facing workers with earnings in this range is represented by segment BD in Figure 6.17, which has a slope equal to that of segment AC .

For earnings above \$16,420, the tax credit was gradually phased out, so that when earnings reached \$40,295, the tax credit was zero. Because after \$16,420 each dollar earned reduced the tax credit by 21 cents, the net wage of EITC recipients was only 79 percent of their market wage (note that the slope of segment DE in Figure 6.17 is flatter than the slope of AC).

Figure 6.17

Earned Income Tax Credit
(Unmarried, Two Children),
2009



Looking closely at Figure 6.17, we can see that EITC recipients will be in one of three “zones”: along AB , along BD , or along DE . The incomes of workers in all three zones are enhanced, which means that all EITC recipients experience an income effect that pushes them in the direction of less work. However, the program creates quite different net wage rates in the zones, and therefore the substitution effect differs across zones.

For workers with earnings below \$12,570, the net wage is greater than the market wage (by 40 percent), so along segment AB , workers experience an increase in the price of leisure. Workers with earnings below \$12,570, then, experience a substitution effect that pushes them in the direction of more work. With an income effect and a substitution effect that push in opposite directions, it is uncertain which effect will dominate. What we can predict, though, is that some of those who would have been out of the labor force in the absence of the EITC program will now decide to seek work (earlier, we discussed the fact that for non-participants in the labor force, the substitution effect dominates).

Segments BD and DE represent two other zones, in which theory predicts that labor supply will *fall*. Along BD , the net wage is equal to the market wage, so the price of leisure in this zone is unchanged while income is enhanced. Workers in this zone experience a pure income effect. Along segment DE , the net wage is actually below the market wage, so in this zone, *both* the income and the substitution effects push in the direction of reduced labor supply.

Using economic theory to analyze labor supply responses induced by the constraint in Figure 6.17, we can come up with two predictions. First, if an EITC

EMPIRICAL STUDY

ESTIMATING THE INCOME EFFECT AMONG LOTTERY WINNERS: THE SEARCH FOR “EXOGENEITY”

Regression analysis, described in Appendix 1A, allows us to analyze the effects of one or more *independent* variables on a *dependent* variable. This statistical procedure is based on an important assumption: that each independent variable is *exogenous* (determined by some outside force and not itself influenced by the dependent variable). That is, we assume that the chain of causation runs from the independent variables to the dependent variable, with no feedback from the dependent variable to those that we assume are independent.

The issue of exogeneity arises when estimating the effects on hours of work caused by a change in income (wages held constant). Theory leads us to predict that desired hours of work are a function of wages, wealth, and preferences. Wealth is not usually observed in most data sets, so *nonlabor* income, such as the returns from financial investments, is used as a proxy for it. Measuring the effect that nonlabor income (an independent, or causal, variable) has on desired hours of work (our dependent variable), holding the wage constant, is intended to capture the income effect predicted by labor supply theory.

The problem is that those who have strong preferences for income and weak preferences for leisure, for example, may tend to accumulate financial assets over time and end up with relatively high levels of nonlabor income later on. Put

differently, high levels of work hours (supposedly our dependent variable) may create high levels of nonlabor income (what we hoped would be our independent variable); thus, when we estimate a correlation between work hours and nonlabor income, we cannot be sure whether we are estimating the income effect, some relationship between hard work and savings, or a mix of both (a problem analogous to the one discussed in the empirical study in chapter 4). In estimating the income effect, therefore, researchers must be careful to use measures of nonlabor income that are truly exogenous and not themselves influenced by the desired hours of work.

Are lottery winnings an exogenous source of nonlabor income? Once a person enters a lottery, winning is a completely random event and thus is not affected by work hours; however, entering the lottery may not be so independent. If those who enter the lottery also have the strongest preferences for leisure, for example, then correlating work hours and lottery winnings across different individuals would not necessarily isolate the income effect. Rather, it might just reflect that those with stronger preferences for leisure (and thus lower work hours) were more likely to enter (and thus win) the lottery.

Therefore, if we want to measure the income effect associated with winning the lottery, we need to find a way to hold both

wages *and* preferences for leisure constant. One study of how winning the lottery affected labor supply took account of the preferences of lottery players by performing a before-and-after analysis using panel data on winners and nonwinners. That is, for winners—defined as receiving prizes over \$20,000, with a median prize of \$635,000—the authors compared hours of work for six years before winning to hours of work during the six years after winning. By focusing on each individual's *changes* in hours and lottery winnings over the two periods, the effects of preferences (which are assumed to be unchanging) drop out of the analysis.

"Nonwinners" in the study were defined as lottery players who won only small prizes, ranging from \$100 to

\$5,000. Labor supply changes for them before and after their small winnings were then calculated and compared to the changes observed among the winners. The study found that for every \$100,000 in prizes, winners reduced their hours of work such that their earnings went down by roughly \$11,000 (that is, winners spent about 11 percent of their prize on "buying" leisure). These findings, of course, are consistent with the predictions concerning the income effect of nonlabor income on labor supply.

Source: Guido W. Imbens, Donald B. Rubin, and Bruce I. Sacerdote, "Estimating the Effect of Unearned Income on Labor Earnings, Savings, and Consumption: Evidence from a Survey of Lottery Players," *American Economic Review* 91 (September 2001): 778–794.

program is started or expanded, we should observe that the labor force participation rate of low-wage workers will increase. Second, a new or expanded EITC program should lead to a reduction in working hours among those along *BD* and *DE* (the effect on hours along *AB* is ambiguous).

Several studies have found evidence consistent with prediction that the EITC should increase labor force participation, with one study finding that over half of the increase in labor force participation among *single mothers* from 1984 to 1996 was caused by expansions in the EITC during that period. The evidence so far, however, does not indicate a measurable drop in hours of work by those receiving the tax credit.²⁶ Thus, the labor supply responses to the EITC are very similar to those found in labor supply studies cited earlier (see footnote 17 and Example 6.5), in that labor force participation rates seem to be more responsive to wage changes than are the hours of work.

²⁶Nada Eissa and Hilary W. Hoynes, "Behavioral Responses to Taxes: Lessons from the EITC and Labor Supply," National Bureau of Economic Research, working paper no. 11729 (November 2005). A study of the labor supply responses to changes in one state's welfare program—which generated a constraint similar to that in Figure 6.17—*did* find the predicted responses in hours of work; see Marianne P. Bitler, Jonah B. Gelbach, and Hilary W. Hoynes, "What Mean Impacts Miss: Distributional Effects of Welfare Reform Experiments," *American Economic Review* 96 (September 2006): 988–1012. For an article that analyzes the effects on wages of the labor-supply responses to the EITC, see Jesse Rothstein, "Is the EITC as Good as an NIT? Conditional Cash Transfers and Tax Incidence," *American Economic Journal: Economic Policy* 2 (February 2010): 177–208.

EXAMPLE 6.7**Wartime Food Requisitions and Agricultural Work Incentives**

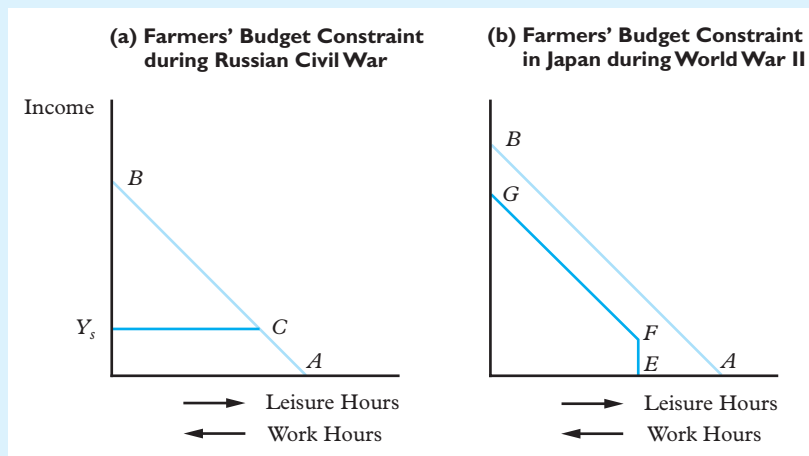
Countries at war often adopt “work requirement” policies to obtain needed food supplies involuntarily from their farming populations. Not surprisingly, the way in which these requisitions are carried out can have enormous effects on the work incentives of farmers. Two alternative methods are contrasted in this example: one was used by the Bolshevik government during the civil war that followed the Russian revolution and the other by Japan during World War II.

From 1917 to 1921, the Bolsheviks requisitioned from farmers all food in excess of the amounts needed for the farmers’ own subsistence; in effect, the surplus was confiscated and given to soldiers and urban dwellers. Graphically, this policy created a budget constraint for farmers like ACY_s in the following diagram (a). Because farmers could keep their output until they reached the subsistence level of income (Y_s), the market wage prevailed until income of Y_s was reached. After that, their net wage was zero (on segment CY_s), because any extra output went to the government. Thus, a prewar market constraint of AB was converted to ACY_s , with the consequence that most farmers maximized

utility near point C . Acreage planted dropped by 27 percent from 1917 to 1921, while harvested output fell by 50 percent!

Japan during World War II handled its food requisitioning policy completely differently. It required a quota to be delivered by each farmer to the government at very low prices, paying farmers the lump sum of EF in diagram (b). Japan, however, allowed farmers to sell any produce above the quota at higher (market) prices. This policy converted the prewar constraint of AB to one much like EFG in diagram (b). In effect, farmers had to work AE hours for the government, for which they were paid EF , but they were then allowed to earn the market wage after that. This policy preserved farmers’ work incentives and apparently created an income effect that increased the total hours of work by Japanese farmers, for despite war-induced shortages of capital and labor, rice production was greater in 1944 than in 1941!

*Data from: Jack Hirshleifer, *Economic Behavior in Adversity* (Chicago: University of Chicago Press, 1987): 16–21, 39–41.*



Review Questions

1. Referring to the definitions in footnote 5, is the following statement true, false, or uncertain? "Leisure must be an inferior good for an individual's labor supply curve to be backward-bending." Explain your answer.
2. Evaluate the following quote: "Higher take-home wages for any group should increase the labor force participation rate for that group."
3. Suppose a government is considering several options to ensure that legal services are provided to the poor:

Option A: All lawyers would be required to devote 5 percent of their work time to the poor, free of charge.

Option B: Lawyers would be required to provide 100 hours of work, free of charge, to the poor.

Option C: Lawyers who earn over \$50,000 in a given year would have to donate \$5,000 to a fund that the government would use to help the poor.

Discuss the likely effects of each option on the hours of work among lawyers. (It would help to *draw* the constraints created by each option.)
4. The way the workers' compensation system works now, employees permanently injured on the job receive a payment of \$X each year, whether they work or not. Suppose the government were to implement a new program in which those who did not work at all got \$0.5X, but those who did work got \$0.5X plus workers' compensation of 50 cents *for every hour worked* (of course, this subsidy would be in addition to the wages paid by their employers). What would be the change in work incentives associated with this change in the way workers' compensation payments were calculated?
5. A firm wants to offer paid sick leave to its workers, but it wants to encourage them not to abuse it by being unnecessarily absent. The firm is considering two options:
 - a. Ten days of paid sick leave per year; any unused leave days at the end of the year are converted to cash at the worker's daily wage rate.
 - b. Ten days of paid sick leave per year; if no sick days are used for two consecutive years, the company agrees to buy the worker a \$100,000 life insurance policy.

Compare the work-incentive effects of the two options, both immediately and in the long run.
6. In 2002, a French law went into effect that cut the standard workweek from 39 to 35 hours (workers got paid for 39 hours even though they worked 35) while at the same time prohibiting overtime hours from being worked. (Overtime in France is paid at 25 percent above the normal wage rate).
 - a. Draw the old budget constraint, showing the overtime premium after 39 hours of work.
 - b. Draw the new budget constraint.
 - c. Analyze which workers in France are better off under the 2002 law. Are any worse off? Explain.
7. Suppose there is a proposal to provide poor people with housing subsidies that are tied to their income levels. These subsidies will be in the form of vouchers the poor can turn over to their landlords in full or partial payment of their housing expenses. The yearly subsidy will equal \$2,400 as long as earnings do not exceed \$8,000 per year. The subsidy is to be reduced 60 cents for every dollar earned in excess of \$8,000; that is, when earnings

reach \$12,000, the person is no longer eligible for rent subsidies.

Draw an arbitrary budget constraint for a person, assuming that he or she receives no government subsidies. Then draw in the budget constraint that arises from the above housing subsidy proposal. After drawing in the budget constraint associated with the proposal, analyze the effects of this proposed housing subsidy program on the labor supply behavior of various groups in the population.

8. The Tax Reform Act of 1986 was designed to reduce the marginal tax rate (the tax rate on the last dollars earned) while eliminating enough deductions and loopholes so that total revenues collected by the government could remain constant. Analyze the work-incentive effects of tax reforms that lower marginal tax rates while keeping total tax revenues constant.
9. The current UI program in the United States gives workers \$X per day if they are unemployed but zero if they take a job for even 1 hour per day. Suppose that the

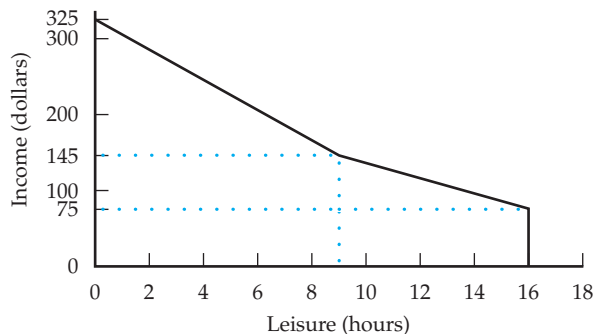
law is changed so that UI beneficiaries can keep getting benefits of \$X per day if they work 2 or fewer hours per day, but if they work more than 2 hours per day, their UI benefits end. Draw the old and new budget constraints (clearly labeled) associated with the UI program, and analyze the work incentives of this proposed change.

10. Assume that the current Disability Insurance (DI) benefit for those who are unable to work is \$X per day and that DI benefits go to *zero* if a worker accepts a job for even 1 hour per week. Suppose that the benefit rules are changed so those disabled workers who take jobs that pay less than \$X per day receive a benefit that brings *their total daily income* (earnings plus the DI benefit) up to \$X. As soon as their labor market earnings rise above \$X per day, their disability benefits end. Draw the old and new budget constraints (label each clearly) associated with the DI program, and analyze the work-incentive effects of the change in benefits.

Problems

1. When the Fair Labor Standards Act began to mandate paying 50 percent more for overtime work, many employers tried to avoid it by cutting hourly pay so that total pay and hours remained the same.
 - a. Assuming that this 50 percent overtime pay premium is newly required for all work beyond eight hours per day, draw a budget constraint that pictures a strategy of cutting hourly pay so that at the original hours of work, total earnings remain the same.
 - b. Suppose that an employer initially paid \$11 per hour and had a 10-hour workday. What hourly base wage will the employer offer so that the total pay for a 10-hour workday will stay the same?
 - c. Will employees who used to work 10 hours per day want to work more or fewer than 10 hours in the new environment (which includes the new wage rate and the mandated overtime premium)?
2. Nina is able to select her weekly work hours. When a new bridge opens up, it cuts one hour off Nina's total daily commute to work. If both leisure and income are normal goods, what is the effect of the shorter commute on Nina's work time?
3. Suppose you win a lottery, and your after-tax gain is \$50,000 per year until

- you retire. As a result, you decide to work part time at 30 hours per week in your old job instead of the usual 40 hours per week.
- Calculate the annual income effect from this lottery gain based on a 50-week year. Interpret the results in light of the theory presented in this chapter.
 - What is the substitution effect associated with this lottery win? Explain.
- The federal minimum wage was increased on July 24, 2007, to \$5.85 from \$5.15. If 16 hours per day are available for work and leisure, draw the daily budget constraint for a worker who was earning the minimum wage rate of \$5.15 and the new budget constraint after the increase.
 - Suppose Michael receives \$50 per day as interest on an inheritance. His wage rate is \$20 per hour, and he can work a maximum of 16 hours per day at his job. Draw his daily budget constraint.
 - Stella can work up to 16 hours per day at her job. Her wage rate is \$8.00 per hour for the first 8 hours. If she works more than 8 hours, her employer pays “time and a half.” Draw Stella’s daily budget constraint.
 - Teddy’s daily budget constraint is shown in the following chart. Teddy’s employer pays him a base wage rate plus overtime if he works more than the standard hours. What is Teddy’s daily nonlabor income? What is Teddy’s base wage rate? What is Teddy’s overtime wage rate? How many hours does Teddy need to work to receive overtime?



Selected Readings

- Blank, Rebecca M. “Evaluating Welfare Reform in the United States.” *Journal of Economic Literature* 40 (December 2002): 1105–1166.
- Card, David E., and Rebecca M. Blank, eds. *Finding Jobs: Work and Welfare Reform*. New York: Russell Sage Foundation, 2000.
- Hoffman, Saul D., and Laurence S. Seidman. *Helping Working Families: The Earned Income Tax Credit*. Kalamazoo, Mich.: W. E. Upjohn Institute for Employment Research, 2002.
- Killingsworth, Mark R. *Labor Supply*. Cambridge, England: Cambridge University Press, 1983.
- Linder, Staffan B. *The Harried Leisure Class*. New York: Columbia University Press, 1970.
- Moffitt, Robert. “Incentive Effects of the U.S. Welfare System: A Review.” *Journal of Economic Literature* 30 (March 1992): 1–62.
- Pencavel, John. “Labor Supply of Men: A Survey.” In *Handbook of Labor Economics*, eds. Orley Ashenfelter and David Card. Amsterdam, N.Y.: Elsevier, 1999.

CHAPTER 7

Uncertainty and Consumer Behavior

So far, we have assumed that prices, incomes, and other variables are known with certainty. However, many of the choices that people make involve considerable uncertainty. Most people, for example, borrow to finance large purchases, such as a house or a college education, and plan to pay for them out of future income. But for most of us, future incomes are uncertain. Our earnings can go up or down; we can be promoted or demoted, or even lose our jobs. And if we delay buying a house or investing in a college education, we risk price increases that could make such purchases less affordable. How should we take these uncertainties into account when making major consumption or investment decisions?

Sometimes we must choose how much *risk* to bear. What, for example, should you do with your savings? Should you invest your money in something safe, such as a savings account, or something riskier but potentially more lucrative, such as the stock market? Another example is the choice of a job or career. Is it better to work for a large, stable company with job security but slim chance for advancement, or is it better to join (or form) a new venture that offers less job security but more opportunity for advancement?

To answer such questions, we must examine the ways that people can compare and choose among risky alternatives. We will do this by taking the following steps:

1. In order to compare the riskiness of alternative choices, we need to quantify risk. We therefore begin this chapter by discussing measures of risk.
2. We will examine people's preferences toward risk. Most people find risk undesirable, but some people find it more undesirable than others.
3. We will see how people can sometimes reduce or eliminate risk. Sometimes risk can be reduced by diversification, by buying insurance, or by investing in additional information.
4. In some situations, people must choose the amount of risk they wish to bear. A good example is investing in stocks or bonds. We will see that such investments involve trade-offs between the monetary gain that one can expect and the riskiness of that gain.
5. Sometimes demand for a good is driven partly or entirely by speculation—people buy the good because they think its price will rise.



CHAPTER OUTLINE

- 7.1 Describing Risk
160
- 7.2 Preferences Toward Risk
165
- 7.3 Reducing Risk
170
- *7.4 The Demand for Risky Assets
176
- 5.5 Bubbles
185
- 7.6 Behavioral Economics
189

LIST OF EXAMPLES

- 7.1 Deterring Crime
164
- 7.2 Business Executives and the Choice of Risk
169
- 7.3 The Value of Title Insurance When Buying a House
173
- 7.4 The Value of Information in an Online Consumer Electronics Market
175
- 7.5 Doctors, Patients, and the Value of Information
175
- 7.6 Investing in the Stock Market
183
- 7.7 The Housing Price Bubble (I)
186
- 7.8 The Housing Price Bubble (II)
188
- 7.9 Selling a House
192
- 7.10 New York City Taxicab Drivers
196



We will see how this can lead to a bubble, where more and more people, convinced that the price will keep rising, buy the good and push its price up further—until eventually the bubble bursts and the price plummets.

In a world of uncertainty, individual behavior may sometimes seem unpredictable, even irrational, and perhaps contrary to the basic assumptions of consumer theory. In the final section of this chapter, we offer an overview of the flourishing field of behavioral economics, which, by introducing important ideas from psychology, has broadened and enriched the study of microeconomics.

7.1 Describing Risk

To describe risk quantitatively, we begin by listing all the possible outcomes of a particular action or event, as well as the likelihood that each outcome will occur.¹ Suppose, for example, that you are considering investing in a company that explores for offshore oil. If the exploration effort is successful, the company's stock will increase from \$30 to \$40 per share; if not, the price will fall to \$20 per share. Thus there are two possible future outcomes: a \$40-per-share price and a \$20-per-share price.

Probability

• **probability** Likelihood that a given outcome will occur.

Probability is the likelihood that a given outcome will occur. In our example, the probability that the oil exploration project will be successful might be $1/4$ and the probability that it is unsuccessful $3/4$. (Note that the probabilities for all possible events must add up to 1.)

Our interpretation of probability can depend on the nature of the uncertain event, on the beliefs of the people involved, or both. One *objective* interpretation of probability relies on the frequency with which certain events tend to occur. Suppose we know that of the last 100 offshore oil explorations, 25 have succeeded and 75 failed. In that case, the probability of success of $1/4$ is objective because it is based directly on the frequency of similar experiences.

But what if there are no similar past experiences to help measure probability? In such instances, objective measures of probability cannot be deduced and more subjective measures are needed. *Subjective probability* is the perception that an outcome will occur. This perception may be based on a person's judgment or experience, but not necessarily on the frequency with which a particular outcome has actually occurred in the past. When probabilities are subjectively determined, different people may attach different probabilities to different outcomes and thereby make different choices. For example, if the search for oil were to take place in an area where no previous searches had ever occurred, I might attach a higher subjective probability than you to the chance that the project will succeed: Perhaps I know more about the project or I have a better understanding of the oil business and can therefore make better use of our common information. Either different information or different abilities to process the same information can cause subjective probabilities to vary among individuals.

¹Some people distinguish between uncertainty and risk along the lines suggested some 60 years ago by economist Frank Knight. *Uncertainty* can refer to situations in which many outcomes are possible but the likelihood of each is unknown. *Risk* then refers to situations in which we can list all possible outcomes and know the likelihood of each occurring. In this chapter, we will always refer to risky situations, but will simplify the discussion by using *uncertainty* and *risk* interchangeably.



Regardless of the interpretation of probability, it is used in calculating two important measures that help us describe and compare risky choices. One measure tells us the *expected value* and the other the *variability* of the possible outcomes.

Expected Value

The **expected value** associated with an uncertain situation is a weighted average of the **payoffs** or values associated with all possible outcomes. The probabilities of each outcome are used as weights. Thus the expected value measures the *central tendency*—the payoff or value that we would expect on average.

Our offshore oil exploration example had two possible outcomes: Success yields a payoff of \$40 per share, failure a payoff of \$20 per share. Denoting “probability of” by Pr , we express the expected value in this case as

$$\begin{aligned}\text{Expected value} &= Pr(\text{success})(\$40/\text{share}) + Pr(\text{failure})(\$20/\text{share}) \\ &= (1/4)(\$40/\text{share}) + (3/4)(\$20/\text{share}) = \$25/\text{share}\end{aligned}$$

More generally, if there are two possible outcomes having payoffs X_1 and X_2 and if the probabilities of each outcome are given by Pr_1 and Pr_2 , then the expected value is

$$E(X) = Pr_1X_1 + Pr_2X_2$$

When there are n possible outcomes, the expected value becomes

$$E(X) = Pr_1X_1 + Pr_2X_2 + \cdots + Pr_nX_n$$

Variability

Variability is the extent to which the possible outcomes of an uncertain situation differ. To see why variability is important, suppose you are choosing between two part-time summer sales jobs that have the same expected income (\$1500). The first job is based entirely on commission—the income earned depends on how much you sell. There are two equally likely payoffs for this job: \$2000 for a successful sales effort and \$1000 for one that is less successful. The second job is salaried. It is very likely (.99 probability) that you will earn \$1510, but there is a .01 probability that the company will go out of business, in which case you would earn only \$510 in severance pay. Table 5.1 summarizes these possible outcomes, their payoffs, and their probabilities.

Note that these two jobs have the same expected income. For Job 1, expected income is $.5(\$2000) + .5(\$1000) = \$1500$; for Job 2, it is $.99(\$1510) + .01(\$510) = \$1500$. However, the *variability* of the possible payoffs is different. We measure

• **expected value** Probability-weighted average of the payoffs associated with all possible outcomes.

• **payoff** Value associated with a possible outcome.

• **variability** Extent to which possible outcomes of an uncertain event differ.

TABLE 7.1 INCOME FROM SALES JOBS

	OUTCOME 1		OUTCOME 2		EXPECTED INCOME (\$)
	PROBABILITY	INCOME (\$)	PROBABILITY	INCOME (\$)	
Job 1: Commission	.5	2000	.5	1000	1500
Job 2: Fixed Salary	.99	1510	.01	510	1500

**TABLE 7.2** DEVIATIONS FROM EXPECTED INCOME (\$)

	OUTCOME 1	DEVIATION	OUTCOME 2	DEVIATION
Job 1	2000	500	1000	–500
Job 2	1510	10	510	–990

• **deviation** Difference between expected payoff and actual payoff.

• **standard deviation** Square root of the weighted average of the squares of the deviations of the payoffs associated with each outcome from their expected values.

variability by recognizing that large differences between actual and expected payoffs (whether positive or negative) imply greater risk. We call these differences **deviations**. Table 5.2 shows the deviations of the possible income from the expected income from each job.

By themselves, deviations do not provide a measure of variability. Why? Because they are sometimes positive and sometimes negative, and as you can see from Table 5.2, the average of the probability-weighted deviations is always 0.² To get around this problem, we square each deviation, yielding numbers that are always positive. We then measure variability by calculating the **standard deviation**: the square root of the average of the *squares* of the deviations of the payoffs associated with each outcome from their expected values.³

Table 5.3 shows the calculation of the standard deviation for our example. Note that the average of the squared deviations under Job 1 is given by

$$.5(\$250,000) + .5(\$250,000) = \$250,000$$

The standard deviation is therefore equal to the square root of \$250,000, or \$500. Likewise, the probability-weighted average of the squared deviations under Job 2 is

$$.99(\$100) + .01(\$980,100) = \$9900$$

The standard deviation is the square root of \$9900, or \$99.50. Thus the second job is much less risky than the first; the standard deviation of the incomes is much lower.⁴

The concept of standard deviation applies equally well when there are many outcomes rather than just two. Suppose, for example, that the first summer job yields incomes ranging from \$1000 to \$2000 in increments of \$100 that are all equally likely. The second job yields incomes from \$1300 to \$1700 (again in increments of \$100) that are also equally likely. Figure 5.1 shows the alternatives

TABLE 7.3 CALCULATING VARIANCE (\$)

	OUTCOME 1	DEVIATION SQUARED	OUTCOME 2	DEVIATION SQUARED	WEIGHTED AVERAGE DEVIATION SQUARED	STANDARD DEVIATION
Job 1	2000	250,000	1000	250,000	250,000	500
Job 2	1510	100	510	980,100	9900	99.50

²For Job 1, the average deviation is $.5(\$500) + .5(-\$500) = 0$; for Job 2 it is $.99(\$10) + .01(-\$990) = 0$.

³Another measure of variability, *variance*, is the square of the standard deviation.

⁴In general, when there are two outcomes with payoffs X_1 and X_2 , occurring with probability Pr_1 and Pr_2 , and $E(X)$ is the expected value of the outcomes, the standard deviation is given by σ , where

$$\sigma^2 = \text{Pr}_1[(X_1 - E(X))^2] + \text{Pr}_2[(X_2 - E(X))^2]$$

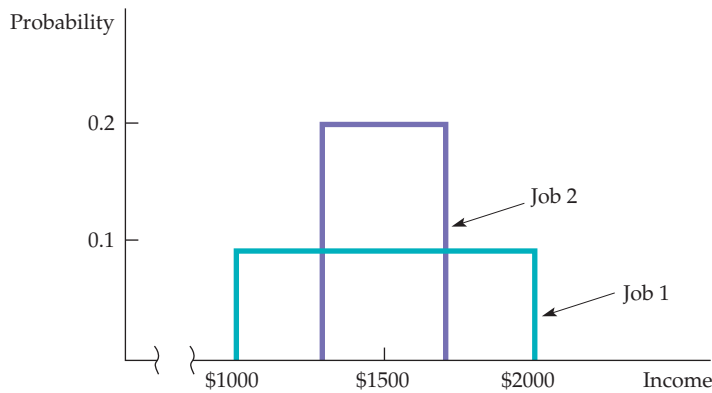


FIGURE 7.1
OUTCOME PROBABILITIES FOR TWO JOBS

The distribution of payoffs associated with Job 1 has a greater spread and a greater standard deviation than the distribution of payoffs associated with Job 2. Both distributions are flat because all outcomes are equally likely.

graphically. (If there had been only two equally probable outcomes, then the figure would be drawn as two vertical lines, each with a height of 0.5.)

You can see from Figure 5.1 that the first job is riskier than the second. The “spread” of possible payoffs for the first job is much greater than the spread for the second. As a result, the standard deviation of the payoffs associated with the first job is greater than that associated with the second.

In this particular example, all payoffs are equally likely. Thus the curves describing the probabilities for each job are flat. In many cases, however, some payoffs are more likely than others. Figure 5.2 shows a situation in which the most extreme payoffs are the least likely. Again, the salary from Job 1 has a greater standard deviation. From this point on, we will use the standard deviation of payoffs to measure the degree of risk.

Decision Making

Suppose you are choosing between the two sales jobs described in our original example. Which job would you take? If you dislike risk, you will take the second job: It offers the same expected income as the first but with less risk. But suppose we add \$100 to each of the payoffs in the first job, so that the expected payoff increases from \$1500 to \$1600. Table 5.4 gives the new earnings and the squared deviations.

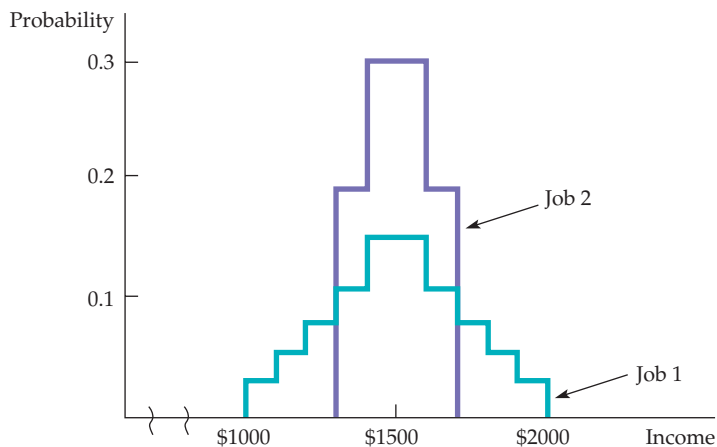


FIGURE 7.2
UNEQUAL PROBABILITY OUTCOMES

The distribution of payoffs associated with Job 1 has a greater spread and a greater standard deviation than the distribution of payoffs associated with Job 2. Both distributions are peaked because the extreme payoffs are less likely than those near the middle of the distribution.

**TABLE 7.4 INCOMES FROM SALES JOBS—MODIFIED (\$)**

	OUTCOME 1	DEVIATION SQUARED	OUTCOME 2	DEVIATION SQUARED	EXPECTED INCOME	STANDARD DEVIATION
Job 1	2100	250,000	1100	250,000	1600	500
Job 2	1510	100	510	980,100	1500	99.50

The two jobs can now be described as follows:

Job 1:	Expected Income = \$1600	Standard Deviation = \$500
Job 2:	Expected Income = \$1500	Standard Deviation = \$99.50

Job 1 offers a higher expected income but is much riskier than Job 2. Which job is preferred depends on the individual. While an aggressive entrepreneur who doesn't mind taking risks might choose Job 1, with the higher expected income and higher standard deviation, a more conservative person might choose the second job.

People's attitudes toward risk affect many of the decisions they make. In Example 5.1 we will see how attitudes toward risk affect people's willingness to break the law, and how this has implications for the fines that should be set for various violations. Then in Section 5.2, we will further develop our theory of consumer choice by examining people's risk preferences in greater detail.

EXAMPLE 7.1 DETERRING CRIME

Fines may be better than incarceration in deterring certain types of crimes, such as speeding, double-parking, tax evasion, and air polluting.⁵ A person choosing to violate the law in these ways has good information and can reasonably be assumed to be behaving rationally.

Other things being equal, the greater the fine, the more a potential criminal will be discouraged from committing the crime. For example, if it cost nothing to catch criminals, and if the crime imposed a calculable cost of \$1000 on society, we might choose to catch all violations and impose a fine of \$1000 on each. This practice would discourage people whose benefit from engaging in the activity was less than the \$1000 fine.

In practice, however, it is very costly to catch law-breakers. Therefore, we save on administrative costs by imposing relatively high fines (which are no more costly to collect than low fines), while allocating

resources so that only a fraction of the violators are apprehended. Thus the size of the fine that must be imposed to discourage criminal behavior depends on the attitudes toward risk of potential violators.

Suppose that a city wants to deter people from double-parking. By double-parking, a typical resident saves \$5 in terms of his own time for engaging in activities that are more pleasant than searching for a parking space. If it costs nothing to catch a double-parker, a fine of just over \$5—say, \$6—should be assessed every time he double-parks. This policy will ensure that the net benefit of double-parking (the \$5 benefit less the \$6 fine) would be less than zero. Our citizen will therefore choose to obey the law. In fact, all potential violators whose benefit was less than or equal to \$5 would be discouraged, while a few whose benefit was greater than \$5 (say, someone who double-parks because of an emergency) would violate the law.

⁵This discussion builds indirectly on Gary S. Becker, "Crime and Punishment: An Economic Approach," *Journal of Political Economy* (March/April 1968): 169–217. See also A. Mitchell Polinsky and Steven Shavell, "The Optimal Tradeoff Between the Probability and the Magnitude of Fines," *American Economic Review* 69 (December 1979): 880–91.



In practice, it is too costly to catch all violators. Fortunately, it's also unnecessary. The same deterrence effect can be obtained by assessing a fine of \$50 and catching only one in ten violators (or perhaps a fine of \$500 with a one-in-100 chance of being caught). In each case, the expected penalty is \$5, i.e., $[\$50][.1]$ or $[\$500][.01]$. A policy that combines a high fine and a low probability of apprehension is likely to reduce enforcement costs. This approach is especially effective if drivers don't like to take risks. In our example, a \$50 fine with a .1 probability of being

caught might discourage most people from violating the law. We will examine attitudes toward risk in the next section.

A new type of crime that has become a serious problem for music and movie producers is digital piracy; it is particularly difficult to catch and fines are rarely imposed. Nevertheless, fines that are levied are often very high. In 2009, a woman was fined \$1.9 million for illegally downloading 24 songs. That amounts to a fine of \$80,000 per song.

7.2 Preferences Toward Risk

We used a job example to show how people might evaluate risky outcomes, but the principles apply equally well to other choices. In this section, we concentrate on consumer choices generally and on the *utility* that consumers obtain from choosing among risky alternatives. To simplify things, we'll consider the utility that a consumer gets from his or her income—or, more appropriately, the market basket that the consumer's income can buy. We now measure payoffs, therefore, in terms of utility rather than dollars.

Figure 5.3 (a) shows how we can describe one woman's preferences toward risk. The curve OE , which gives her utility function, tells us the level of utility (on the vertical axis) that she can attain for each level of income (measured in thousands of dollars on the horizontal axis). The level of utility increases from 10 to 16 to 18 as income increases from \$10,000 to \$20,000 to \$30,000. But note that *marginal utility* is diminishing, falling from 10 when income increases from 0 to \$10,000, to 6 when income increases from \$10,000 to \$20,000, and to 2 when income increases from \$20,000 to \$30,000.

Now suppose that our consumer has an income of \$15,000 and is considering a new but risky sales job that will either double her income to \$30,000 or cause it to fall to \$10,000. Each possibility has a probability of .5. As Figure 5.3 (a) shows, the utility level associated with an income of \$10,000 is 10 (at point A) and the utility level associated with an income of \$30,000 is 18 (at E). The risky job must be compared with the current \$15,000 job, for which the utility is 13.5 (at B).

To evaluate the new job, she can calculate the expected value of the resulting income. Because we are measuring value in terms of her utility, we must calculate the **expected utility** $E(u)$ that she can obtain. The expected utility is *the sum of the utilities associated with all possible outcomes, weighted by the probability that each outcome will occur*. In this case expected utility is

$$E(u) = (1/2)u(\$10,000) + (1/2)u(\$30,000) = 0.5(10) + 0.5(18) = 14$$

The risky new job is thus preferred to the original job because the expected utility of 14 is greater than the original utility of 13.5.

The old job involved no risk—it guaranteed an income of \$15,000 and a utility level of 13.5. The new job is risky but offers both a higher expected income (\$20,000) and, more importantly, a higher expected utility. If the woman wishes to increase her expected utility, she will take the risky job.

In §3.1, we explained that a utility function assigns a level of utility to each possible market basket.

In §3.5, marginal utility is described as the additional satisfaction obtained by consuming an additional amount of a good.

• **expected utility** Sum of the utilities associated with all possible outcomes, weighted by the probability that each outcome will occur.



• **risk averse** Condition of preferring a certain income to a risky income with the same expected value.

• **risk neutral** Condition of being indifferent between a certain income and an uncertain income with the same expected value.

• **risk loving** Condition of preferring a risky income to a certain income with the same expected value.

• **risk premium** Maximum amount of money that a risk-averse person will pay to avoid taking a risk.

Different Preferences Toward Risk

People differ in their willingness to bear risk. Some are risk averse, some risk loving, and some risk neutral. An individual who is **risk averse** prefers a certain given income to a risky income with the same expected value. (Such a person has a diminishing marginal utility of income.) Risk aversion is the most common attitude toward risk. To see that most people are risk averse most of the time, note that most people not only buy life insurance, health insurance, and car insurance, but also seek occupations with relatively stable wages.

Figure 5.3 (a) applies to a woman who is risk averse. Suppose hypothetically that she can have either a certain income of \$20,000, or a job yielding an income of \$30,000 with probability .5 and an income of \$10,000 with probability .5 (so that the expected income is also \$20,000). As we saw, the expected utility of the uncertain income is 14—an average of the utility at point A(10) and the utility at E(18)—and is shown by F. Now we can compare the expected utility associated with the risky job to the utility generated if \$20,000 were earned without risk. This latter utility level, 16, is given by D in Figure 5.3 (a). It is clearly greater than the expected utility of 14 associated with the risky job.

For a risk-averse person, losses are more important (in terms of the change in utility) than gains. Again, this can be seen from Figure 5.3 (a). A \$10,000 increase in income, from \$20,000 to \$30,000, generates an increase in utility of two units; a \$10,000 decrease in income, from \$20,000 to \$10,000, creates a loss of utility of six units.

A person who is **risk neutral** is indifferent between a certain income and an uncertain income with the same expected value. In Figure 5.3 (c) the utility associated with a job generating an income of either \$10,000 or \$30,000 with equal probability is 12, as is the utility of receiving a certain income of \$20,000. As you can see from the figure, the marginal utility of income is constant for a risk-neutral person.⁶

Finally, an individual who is **risk loving** prefers an uncertain income to a certain one, even if the expected value of the uncertain income is less than that of the certain income. Figure 5.3 (b) shows this third possibility. In this case, the expected utility of an uncertain income, which will be either \$10,000 with probability .5 or \$30,000 with probability .5, is *higher* than the utility associated with a certain income of \$20,000. Numerically,

$$E(u) = .5u(\$10,000) + .5u(\$30,000) = .5(3) + .5(18) = 10.5 > u(\$20,000) = 8$$

Of course, some people may be averse to some risks and act like risk lovers with respect to others. For example, many people purchase life insurance and are conservative with respect to their choice of jobs, but still enjoy gambling. Some criminologists might describe criminals as risk lovers, especially if they commit crimes despite a high prospect of apprehension and punishment. Except for such special cases, however, few people are risk loving, at least with respect to major purchases or large amounts of income or wealth.

RISK PREMIUM The **risk premium** is the maximum amount of money that a risk-averse person will pay to avoid taking a risk. In general, the magnitude

⁶Thus, when people are risk neutral, the income they earn can be used as an indicator of well-being. A government policy that doubles incomes would then also double their utility. At the same time, government policies that alter the risks that people face, without changing their expected incomes, would not affect their well-being. Risk neutrality allows a person to avoid the complications that might be associated with the effects of governmental actions on the riskiness of outcomes.



of the risk premium depends on the risky alternatives that the person faces. To determine the risk premium, we have reproduced the utility function of Figure 5.3 (a) in Figure 5.4 and extended it to an income of \$40,000. Recall that an expected utility of 14 is achieved by a woman who is going to take a risky job with an expected income of \$20,000. This outcome is shown graphically by drawing a horizontal line to the vertical axis from point *F*, which bisects straight

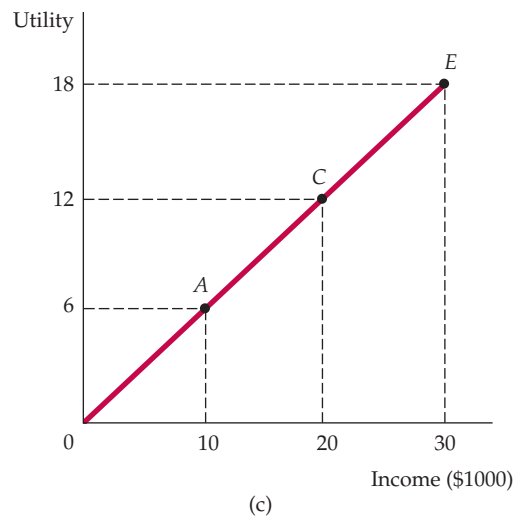
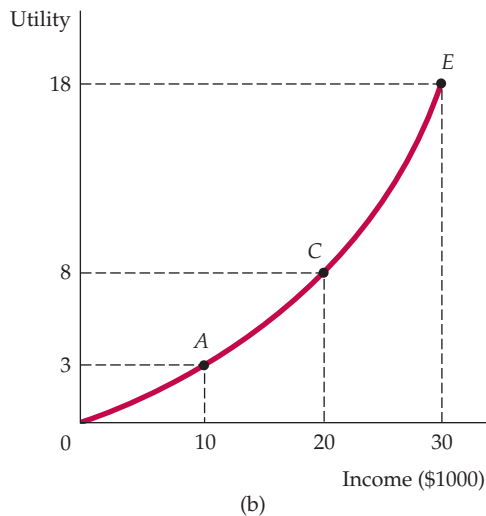
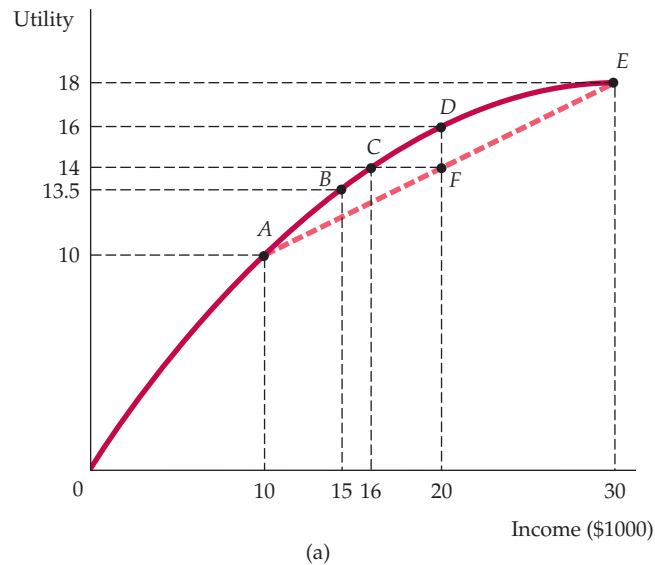


FIGURE 7.3
RISK AVERSE, RISK LOVING, AND RISK NEUTRAL

People differ in their preferences toward risk. In (a), a consumer's marginal utility diminishes as income increases. The consumer is risk averse because she would prefer a certain income of \$20,000 (with a utility of 16) to a gamble with a .5 probability of \$10,000 and a .5 probability of \$30,000 (and expected utility of 14). In (b), the consumer is risk loving: She would prefer the same gamble (with expected utility of 10.5) to the certain income (with a utility of 8). Finally, the consumer in (c) is risk neutral and indifferent between certain and uncertain events with the same expected income.

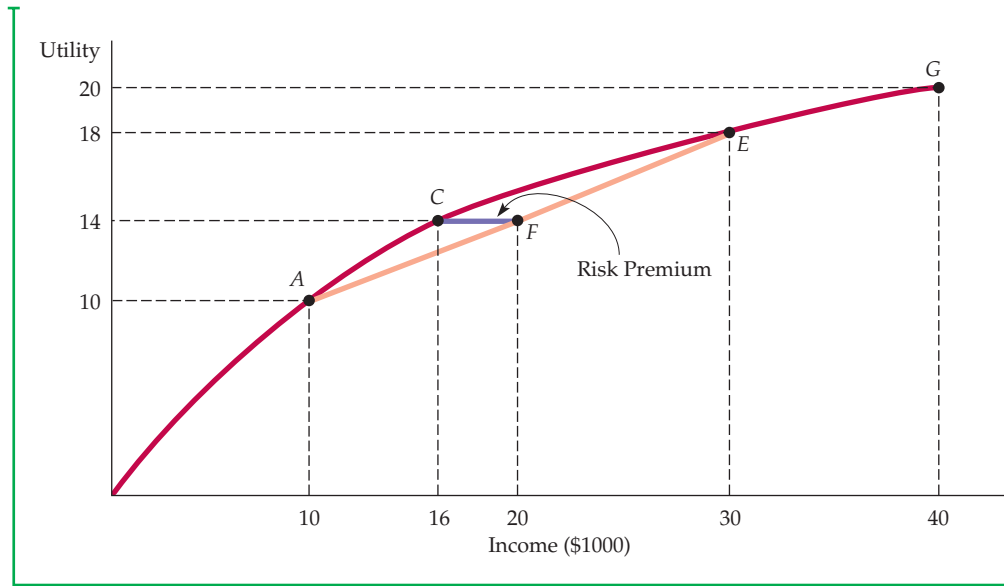


FIGURE 7.4
RISK PREMIUM

The risk premium, CF , measures the amount of income that an individual would give up to leave her indifferent between a risky choice and a certain one. Here, the risk premium is \$4000 because a certain income of \$16,000 (at point C) gives her the same expected utility (14) as the uncertain income (a .5 probability of being at point A and a .5 probability of being at point E) that has an expected value of \$20,000.

line AE (thus representing an average of \$10,000 and \$30,000). But the utility level of 14 can also be achieved if the woman has a *certain* income of \$16,000, as shown by dropping a vertical line from point C . Thus, the risk premium of \$4000, given by line segment CF , is the amount of expected income (\$20,000 minus \$16,000) that she would give up in order to remain indifferent between the risky job and a hypothetical job that would pay her a certain income of \$16,000.

RISK AVERSION AND INCOME The extent of an individual's risk aversion depends on the nature of the risk and on the person's income. Other things being equal, risk-averse people prefer a smaller variability of outcomes. We saw that when there are two outcomes—an income of \$10,000 and an income of \$30,000—the risk premium is \$4000. Now consider a second risky job, also illustrated in Figure 5.4. With this job, there is a .5 probability of receiving an income of \$40,000, with a utility level of 20, and a .5 probability of getting an income of \$0, with a utility level of 0. The expected income is again \$20,000, but the expected utility is only 10:

$$\text{Expected utility} = .5u(\$0) + .5u(\$40,000) = 0 + .5(20) = 10$$

Compared to a hypothetical job that pays \$20,000 with certainty, the person holding this risky job gets 6 fewer units of expected utility: 10 rather than 16 units. At the same time, however, this person could also get 10 units of utility from a job that pays \$10,000 with certainty. Thus the risk premium in this case is \$10,000, because this person would be willing to give up \$10,000 of her \$20,000 expected income to avoid bearing the risk of an uncertain income. The greater the variability of income, the more the person would be willing to pay to avoid the risky situation.

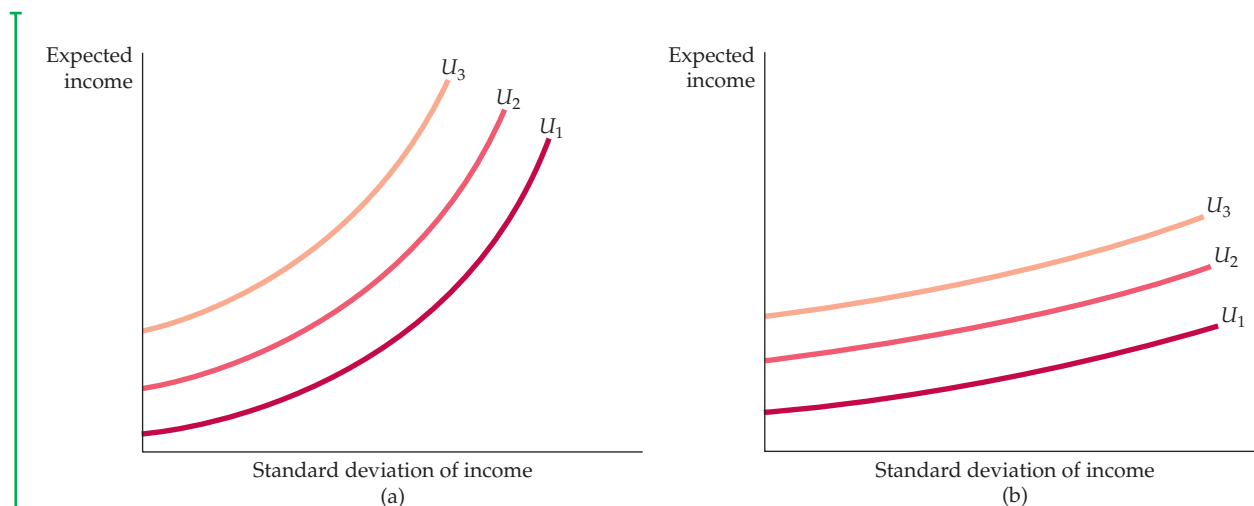


FIGURE 7.5
RISK AVERSION AND INDIFFERENCE CURVES

Part **(a)** applies to a person who is highly risk averse: An increase in this individual's standard deviation of income requires a large increase in expected income if he or she is to remain equally well off. Part **(b)** applies to a person who is only slightly risk averse: An increase in the standard deviation of income requires only a small increase in expected income if he or she is to remain equally well off.

RISK AVERSION AND INDIFFERENCE CURVES We can also describe the extent of a person's risk aversion in terms of indifference curves that relate expected income to the variability of income, where the latter is measured by the standard deviation. Figure 5.5 shows such indifference curves for two individuals, one who is highly risk averse and another who is only slightly risk averse. Each indifference curve shows the combinations of expected income and standard deviation of income that give the individual the same amount of utility. Observe that all of the indifference curves are upward sloping: Because risk is undesirable, the greater the amount of risk, the greater the expected income needed to make the individual equally well off.

Figure 5.5 (a) describes an individual who is highly risk averse. Observe that in order to leave this person equally well off, an increase in the standard deviation of income requires a large increase in expected income. Figure 5.5 (b) applies to a slightly risk-averse person. In this case, a large increase in the standard deviation of income requires only a small increase in expected income.

In §3.1, we define an indifference curve as all market baskets that generate the same level of satisfaction for a consumer.

EXAMPLE 7.2 BUSINESS EXECUTIVES AND THE CHOICE OF RISK

Are business executives more risk loving than most people? When they are presented with alternative strategies, some risky, some safe, which do they choose? In one study, 464 executives were asked

to respond to a questionnaire describing risky situations that an individual might face as vice president of a hypothetical company.⁷ Respondents were presented with four risky events, each of which had a

⁷This example is based on Kenneth R. MacCrimmon and Donald A. Wehrung, "The Risk In-Basket," *Journal of Business* 57 (1984): 367–87.



given probability of a favorable and unfavorable outcome. The payoffs and probabilities were chosen so that each event had the same expected value. In increasing order of the risk involved (as measured by the difference between the favorable and unfavorable outcomes), the four items were:

1. A lawsuit involving a patent violation
2. A customer threatening to buy from a competitor
3. A union dispute
4. A joint venture with a competitor

To gauge their willingness to take or avoid risks, researchers asked respondents a series of questions regarding business strategy. In one situation, they could pursue a risky strategy with the possibility of a high return right away or delay making a choice until the outcomes became more certain and the risk was reduced. In another situation, respondents could opt for an immediately risky but potentially profitable strategy that could lead to a promotion, or they could delegate the decision to someone else, which

would protect their job but eliminate the promotion possibility.

The study found that executives vary substantially in their preferences toward risk. Roughly 20 percent indicated that they were relatively risk neutral; 40 percent opted for the more risky alternatives; and 20 percent were clearly risk averse (20 percent did not respond). More importantly, executives (including those who chose risky alternatives) typically made efforts to reduce or eliminate risk, usually by delaying decisions and collecting more information.

Some have argued that a cause of the financial crisis of 2008 was excessive risk-taking by bankers and Wall Street executives who could earn huge bonuses if their ventures succeeded but faced very little downside if the ventures failed. The U.S. Treasury Department's Troubled Asset Relief Program (TARP) bailed out some of the banks, but so far has been unable to impose constraints on "unnecessary and excessive" risk-taking by banks' executives.

We will return to the use of indifference curves as a means of describing risk aversion in Section 5.4, where we discuss the demand for risky assets. First, however, we will turn to the ways in which an individual can reduce risk.

7.3 Reducing Risk

As the recent growth in state lotteries shows, people sometimes choose risky alternatives that suggest risk-loving rather than risk-averse behavior. Most people, however, spend relatively small amounts on lottery tickets and casinos. When more important decisions are involved, they are generally risk averse. In this section, we describe three ways by which both consumers and businesses commonly reduce risks: *diversification*, *insurance*, and *obtaining more information* about choices and payoffs.

Diversification

Recall the old saying, "Don't put all your eggs in one basket." Ignoring this advice is unnecessarily risky: If your basket turns out to be a bad bet, all will be lost. Instead, you can reduce risk through **diversification**: allocating your resources to a variety of activities whose outcomes are not closely related.

Suppose, for example, that you plan to take a part-time job selling appliances on a commission basis. You can decide to sell only air conditioners or only heaters, or you can spend half your time selling each. Of course, you can't be sure how hot or cold the weather will be next year. How should you apportion your time in order to minimize the risk involved?

Risk can be minimized by *diversification*—by allocating your time so that you sell two or more products (whose sales are not closely related) rather than a

• **diversification** Practice of reducing risk by allocating resources to a variety of activities whose outcomes are not closely related.

**TABLE 7.5** INCOME FROM SALES OF APPLIANCES (\$)

	HOT WEATHER	COLD WEATHER
Air conditioner sales	30,000	12,000
Heater sales	12,000	30,000

single product. Suppose there is a 0.5 probability that it will be a relatively hot year, and a 0.5 probability that it will be cold. Table 5.5 gives the earnings that you can make selling air conditioners and heaters.

If you sell only air conditioners or only heaters, your actual income will be either \$12,000 or \$30,000, but your expected income will be \$21,000 ($.5[\$30,000] + .5[\$12,000]$). But suppose you diversify by dividing your time evenly between the two products. In that case, your income will certainly be \$21,000, regardless of the weather. If the weather is hot, you will earn \$15,000 from air conditioner sales and \$6,000 from heater sales; if it is cold, you will earn \$6,000 from air conditioners and \$15,000 from heaters. In this instance, diversification eliminates all risk.

Of course, diversification is not always this easy. In our example, heater and air conditioner sales are **negatively correlated variables**—they tend to move in opposite directions; whenever sales of one are strong, sales of the other are weak. But the principle of diversification is a general one: As long as you can allocate your resources toward a variety of activities whose outcomes are *not* closely related, you can eliminate some risk.

• **negatively correlated variables** Variables having a tendency to move in opposite directions.

THE STOCK MARKET Diversification is especially important for people who invest in the stock market. On any given day, the price of an individual stock can go up or down by a large amount, but some stocks rise in price while others fall. An individual who invests all her money in a single stock (i.e., puts all her eggs in one basket) is therefore taking much more risk than necessary. Risk can be reduced—although not eliminated—by investing in a portfolio of ten or twenty different stocks. Likewise, you can diversify by buying shares in **mutual funds**: organizations that pool funds of individual investors to buy a large number of different stocks. There are thousands of mutual funds available today for both stocks and bonds. These funds are popular because they reduce risk through diversification and because their fees are typically much lower than the cost of assembling one's own portfolio of stocks.

• **mutual fund** Organization that pools funds of individual investors to buy a large number of different stocks or other financial assets.

In the case of the stock market, not all risk is diversifiable. Although some stocks go up in price when others go down, stock prices are to some extent **positively correlated variables**: They tend to move in the same direction in response to changes in economic conditions. For example, the onset of a severe recession, which is likely to reduce the profits of many companies, may be accompanied by a decline in the overall market. Even with a diversified portfolio of stocks, therefore, you still face some risk.

• **positively correlated variables** Variables having a tendency to move in the same direction.

Insurance

We have seen that risk-averse people are willing to pay to avoid risk. In fact, if the cost of insurance is equal to the expected loss (e.g., a policy with an expected loss of \$1,000 will cost \$1,000), risk-averse people will buy enough insurance to recover fully from any financial losses they might suffer.



Why? The answer is implicit in our discussion of risk aversion. Buying insurance assures a person of having the same income whether or not there is a loss. Because the insurance cost is equal to the expected loss, this certain income is equal to the expected income from the risky situation. For a risk-averse consumer, the guarantee of the same income regardless of the outcome generates more utility than would be the case if that person had a high income when there was no loss and a low income when a loss occurred.

To clarify this point, let's suppose a homeowner faces a 10-percent probability that his house will be burglarized and he will suffer a \$10,000 loss. Let's assume he has \$50,000 worth of property. Table 5.6 shows his wealth in two situations—with insurance costing \$1000 and without insurance.

Note that expected wealth is the same (\$49,000) in both situations. The variability, however, is quite different. As the table shows, with no insurance the standard deviation of wealth is \$3000; with insurance, it is 0. If there is no burglary, the uninsured homeowner gains \$1000 relative to the insured homeowner. But with a burglary, the uninsured homeowner loses \$9000 relative to the insured homeowner. Remember: for a risk-averse individual, losses count more (in terms of changes in utility) than gains. A risk-averse homeowner, therefore, will enjoy higher utility by purchasing insurance.

THE LAW OF LARGE NUMBERS Consumers usually buy insurance from companies that specialize in selling it. Insurance companies are firms that offer insurance because they know that when they sell a large number of policies, they face relatively little risk. The ability to avoid risk by operating on a large scale is based on the *law of large numbers*, which tells us that although single events may be random and largely unpredictable, the average outcome of many similar events can be predicted. For example, I may not be able to predict whether a coin toss will come out heads or tails, but I know that when many coins are flipped, approximately half will turn up heads and half tails. Likewise, if I am selling automobile insurance, I cannot predict whether a particular driver will have an accident, but I can be reasonably sure, judging from past experience, what fraction of a large group of drivers will have accidents.

ACTUARIAL FAIRNESS By operating on a large scale, insurance companies can be sure that over a sufficiently large number of events, total premiums paid in will be equal to the total amount of money paid out. Let's return to our burglary example. A man knows that there is a 10-percent probability that his house will be burgled; if it is, he will suffer a \$10,000 loss. Prior to facing this risk, he calculates the expected loss to be \$1000 ($.10 \times \$10,000$). The risk involved is considerable, however, because there is a 10-percent probability of

TABLE 7.6 THE DECISION TO INSURE (\$)

INSURANCE	BURGLARY (PR = .1)	NO BURGLARY (PR = .9)	EXPECTED WEALTH	STANDARD DEVIATION
No	40,000	50,000	49,000	3000
Yes	49,000	49,000	49,000	0



a large loss. Now suppose that 100 people are similarly situated and that all of them buy burglary insurance from the same company. Because they all face a 10-percent probability of a \$10,000 loss, the insurance company might charge each of them a premium of \$1000. This \$1000 premium generates an insurance fund of \$100,000 from which losses can be paid. The insurance company can rely on the law of large numbers, which holds that the expected loss to the 100 individuals as a whole is likely to be very close to \$1000 each. The total payout, therefore, will be close to \$100,000, and the company need not worry about losing more than that.

When the insurance premium is equal to the expected payout, as in the example above, we say that the insurance is **actuarially fair**. But because they must cover administrative costs and make some profit, insurance companies typically charge premiums *above* expected losses. If there are a sufficient number of insurance companies to make the market competitive, these premiums will be close to actuarially fair levels. In some states, however, insurance premiums are regulated in order to protect consumers from “excessive” premiums. We will examine government regulation of markets in detail in Chapters 9 and 10 of this book.

In recent years, some insurance companies have come to the view that catastrophic disasters such as earthquakes are so unique and unpredictable that they cannot be viewed as diversifiable risks. Indeed, as a result of losses from past disasters, these companies do not feel that they can determine actuarially fair insurance rates. In California, for example, the state itself has had to enter the insurance business to fill the gap created when private companies refused to sell earthquake insurance. The state-run pool offers less insurance coverage at higher rates than was previously offered by private insurers.

- **actuarially fair**

Characterizing a situation in which an insurance premium is equal to the expected payout.

EXAMPLE 7.3 THE VALUE OF TITLE INSURANCE WHEN BUYING A HOUSE

Suppose you are buying your first house. To close the sale, you will need a deed that gives you clear “title.” Without such a clear title, there is always a chance that the seller of the house is not its true owner. Of course, the seller could be engaging in fraud, but it is more likely that the seller is unaware of the exact nature of his or her ownership rights. For example, the owner may have borrowed heavily, using the house as “collateral” for a loan. Or the property might carry with it a legal requirement that limits the use to which it may be put.

Suppose you are willing to pay \$300,000 for the house, but you believe there is a one-in-twenty chance that careful research will reveal that the seller does not actually own the property. The property



would then be worth nothing. If there were no insurance available, a risk-neutral person would bid at most \$285,000 for the property ($.95[\$300,000] + .05[0]$). However, if you expect to tie up most of your assets in the house, you would probably be risk averse and, therefore, bid much less—say, \$230,000.

In situations such as this, it is clearly in the interest of the buyer to be sure that there is no risk of a lack of full ownership. The buyer does this by purchasing “title insurance.” The title insurance company researches the history of the property, checks to see whether any legal liabilities are attached to it, and generally assures itself that there is no ownership problem. The insurance company then agrees to bear any remaining risk that might exist.



Because the title insurance company is a specialist in such insurance and can collect the relevant information relatively easily, the cost of title insurance is often less than the expected value of the loss involved. A fee of \$1500 for title insurance is not unusual, even though the expected loss can be much higher. It is also in the interest of sellers to provide title insurance, because

all but the most risk-loving buyers will pay much more for a house when it is insured than when it is not. In fact, most states require sellers to provide title insurance before a sale can be completed. In addition, because mortgage lenders are all concerned about such risks, they usually require new buyers to have title insurance before issuing a mortgage.

• **value of complete information** Difference between the expected value of a choice when there is complete information and the expected value when information is incomplete.

The Value of Information

People often make decisions based on limited information. If more information were available, one could make better predictions and reduce risk. Because information is a valuable commodity, people will pay for it. The **value of complete information** is the difference between the expected value of a choice when there is complete information and the expected value when information is incomplete.

To see how information can be valuable, suppose you manage a clothing store and must decide how many suits to order for the fall season. If you order 100 suits, your cost is \$180 per suit. If you order only 50 suits, your cost increases to \$200. You know that you will be selling suits for \$300 each, but you are not sure how many you can sell. All suits not sold can be returned, but for only half of what you paid for them. Without additional information, you will act on your belief that there is a .5 probability that you will sell 100 suits and a .5 probability that you will sell 50. Table 5.7 gives the profit that you would earn in each of these two cases.

Without additional information, you would choose to buy 100 suits if you were risk neutral, taking the chance that your profit might be either \$12,000 or \$1500. But if you were risk averse, you might buy 50 suits: In that case, you would know for sure that your profit would be \$5000.

With complete information, you can place the correct order regardless of future sales. If sales were going to be 50 and you ordered 50 suits, your profits would be \$5000. If, on the other hand, sales were going to be 100 and you ordered 100 suits, your profits would be \$12,000. Because both outcomes are equally likely, your expected profit with complete information would be \$8500. The value of information is computed as

	Expected value with complete information:	\$8500
Less:	Expected value with uncertainty (buy 100 suits):	<u>−6750</u>
Equals:	Value of complete information	\$1750

Thus it is worth paying up to \$1750 to obtain an accurate prediction of sales. Even though forecasting is inevitably imperfect, it may be worth investing in a marketing study that provides a reasonable forecast of next year's sales.

TABLE 7.7 PROFITS FROM SALES OF SUITS (\$)

	SALES OF 50	SALES OF 100	EXPECTED PROFIT
Buy 50 suits	5000	5000	5000
Buy 100 suits	1500	12,000	6750



EXAMPLE 7.4 THE VALUE OF INFORMATION IN AN ONLINE CONSUMER ELECTRONICS MARKET

Internet-based price comparison sites offer a valuable informational resource to consumers, as shown by a study of a leading price-comparison website, Shopper.com. Researchers studied price information provided to consumers on over 1,000 top-selling electronics products for an 8-month period. They found that consumers saved about 16% when using this website versus shopping in the store, because the website significantly reduced the cost of finding the lowest priced product.⁸

The value of price comparison information is not the same for everyone and for every product. Competition matters. The study found that when only two firms list prices on Shopper.com,

consumers save 11%. But the savings increase with the number of competitors, jumping to 20% when more than 30 companies list prices.

One might think that the Internet will generate so much information about prices that only the lowest-price products will be sold in the long run, causing the value of such information to eventually decline to zero. So far, this has not been the case. There are fixed costs for parties to both transmit and to acquire information over the Internet. These include the costs of maintaining servers and the fees that sites such as Shopper.com charge to list prices at their sites. The result is that prices are likely to continue to vary widely as the Internet continues to grow and mature.

You might think that more information is always a good thing. As the following example shows, however, that is not always the case.

EXAMPLE 7.5 DOCTORS, PATIENTS, AND THE VALUE OF INFORMATION

Suppose you were seriously ill and required major surgery. Assuming you wanted to get the best care possible, how would you go about choosing a surgeon and a hospital to provide that care? Many people would ask their friends or their primary-care physician for a recommendation. Although this might be helpful, a truly informed decision would probably require more detailed information. For example, how successful has a recommended surgeon and her affiliated hospital been in performing the particular operation that you need? How many of her patients have died or had serious complications from the operation, and how do these numbers compare with those for other surgeons and hospitals? This kind of information is likely to be difficult or impossible for most patients to obtain. Would patients be better off if detailed information about the performance records of doctors and hospitals were readily available?



Not necessarily. More information is often, but not always, better. Interestingly in this case, access to performance information could actually lead to worse health outcomes. Why? Because access to such information would create two different incentives that would affect the behavior of both doctors and patients. First, it would allow patients to choose doctors with better performance records, which creates an incentive for doctors to perform better. That is a good thing. But second, it would encourage doctors to limit their practices to patients who are in relatively good health. The reason is that very old or very sick patients are more likely to have complications or die as a result of treatment; doctors who treat such patients are likely to have worse performance records (other factors being equal). To the extent that doctors would be judged according to performance, they would

⁸Michael Baye, John Morgan, and Patrick Scholten, "The Value of Information in an Online Electronics Market," *Journal of Public Policy and Marketing*, vol. 22 (2003): 17–25.



have an incentive to avoid treating very old or sick patients. As a result, such patients would find it difficult or impossible to obtain treatment.

Whether more information is better depends on which effect dominates—the ability of patients to make more informed choices versus the incentive for doctors to avoid very sick patients. In a recent study, economists examined the impact of the mandatory “report cards” introduced in New York and Pennsylvania in the early 1990s to evaluate outcomes of coronary bypass surgeries.⁹ They analyzed hospital choices and outcomes for all elderly heart attack patients and patients receiving coronary bypass surgery in the United States from 1987 through 1994. By comparing trends in New York and Pennsylvania to the trends in other states, they could determine the effect of the increased information made possible by the availability of report cards. They found that although report cards improved matching of patients with

hospitals and doctors, they also caused a shift in treatment from sicker patients towards healthier ones. Overall, this led to worse outcomes, especially among sicker patients. Thus the study concluded that report cards reduced welfare.

The medical profession has responded to this problem to some extent. For example, in 2010, cardiac surgery programs across the country voluntarily reported the results of coronary-artery bypass grafting procedures. Each program was rated with one to three stars, but this time the ratings were “risk adjusted” to reduce the incentive for doctors to choose less risky patients.

More information often improves welfare because it allows people to reduce risk and to take actions that might reduce the effect of bad outcomes. However, as this example makes clear, information can cause people to change their behavior in undesirable ways. We will discuss this issue further in Chapter 17.

*7.4 The Demand for Risky Assets

Most people are risk averse. Given a choice, they prefer fixed monthly incomes to those which, though equally large on average, fluctuate randomly from month to month. Yet many of these same people will invest all or part of their savings in stocks, bonds, and other assets that carry some risk. Why do risk-averse people invest in the stock market and thereby risk losing part or all of their investments?¹⁰ How do people decide how much risk to bear when making investments and planning for the future? To answer these questions, we must examine the demand for risky assets.

Assets

An **asset** is *something that provides a flow of money or services to its owner*. A home, an apartment building, a savings account, or shares of General Motors stock are all assets. A home, for example, provides a flow of housing services to its owner, and, if the owner did not wish to live there, could be rented out, thereby providing a monetary flow. Likewise, apartments can be rented out, providing a flow of rental income to the owner of the building. A savings account pays interest (usually every day or every month), which is usually reinvested in the account.

• **asset** Something that provides a flow of money or services to its owner.

⁹David Dranove, Daniel Kessler, Mark McClennan, and Mark Satterthwaite, “Is More Information Better? The Effects of ‘Report Cards’ on Health Care Providers,” *Journal of Political Economy* 3 (June 2003): 555–558.

¹⁰Most Americans have at least some money invested in stocks or other risky assets, though often indirectly. For example, many people who hold full-time jobs have shares in pension funds underwritten in part by their own salary contributions and in part by employer contributions. Usually such funds are partly invested in the stock market.



The monetary flow that one receives from asset ownership can take the form of an explicit payment, such as the rental income from an apartment building: Every month, the landlord receives rent checks from the tenants. Another form of explicit payment is the dividend on shares of common stock: Every three months, the owner of a share of General Motors stock receives a quarterly dividend payment.

But sometimes the monetary flow from ownership of an asset is implicit: It takes the form of an increase or decrease in the price or value of the asset. An increase in the value of an asset is a *capital gain*; a decrease is a *capital loss*. For example, as the population of a city grows, the value of an apartment building may increase. The owner of the building will then earn a capital gain beyond the rental income. The capital gain is *unrealized* until the building is sold because no money is actually received until then. There is, however, an implicit monetary flow because the building *could* be sold at any time. The monetary flow from owning General Motors stock is also partly implicit. The price of the stock changes from day to day, and each time it does, owners gain or lose.

Risky and Riskless Assets

A **risky asset** provides a monetary flow that is at least in part random. In other words, the monetary flow is not known with certainty in advance. A share of General Motors stock is an obvious example of a risky asset: You cannot know whether the price of the stock will rise or fall over time, nor can you even be sure that the company will continue to pay the same (or any) dividend per share. Although people often associate risk with the stock market, most other assets are also risky.

An apartment building is one example. You cannot know how much land values will rise or fall, whether the building will be fully rented all the time, or even whether the tenants will pay their rents promptly. Corporate bonds are another example—the issuing corporation could go bankrupt and fail to pay bond owners their interest and principal. Even long-term U.S. government bonds that mature in 10 or 20 years are risky. Although it is highly unlikely that the federal government will go bankrupt, the rate of inflation could unexpectedly increase and make future interest payments and the eventual repayment of principal worth less in real terms, thereby reducing the value of the bonds.

In contrast, a **riskless (or risk-free) asset** pays a monetary flow that is known with certainty. Short-term U.S. government bonds—called Treasury bills—are riskless, or almost riskless. Because they mature in a few months, there is very little risk from an unexpected increase in the rate of inflation. You can also be reasonably confident that the U.S. government will not default on the bond (i.e., refuse to pay back the holder when the bond comes due). Other examples of riskless or almost riskless assets include passbook savings accounts and short-term certificates of deposit.

- **risky asset** Asset that provides an uncertain flow of money or services to its owner.

- **riskless (or risk-free) asset** Asset that provides a flow of money or services that is known with certainty.

Asset Returns

People buy and hold assets because of the monetary flows they provide. To compare assets with each other, it helps to think of this monetary flow relative to an asset's price or value. The **return** on an asset is the total monetary flow it yields—including capital gains or losses—as a fraction of its price. For example, a bond worth \$1000 today that pays out \$100 this year (and every year) has a return of

- **return** Total monetary flow of an asset as a fraction of its price.



10 percent.¹¹ If an apartment building was worth \$10 million last year, increased in value to \$11 million this year, and also provided rental income (after expenses) of \$0.5 million, it would have yielded a return of 15 percent over the past year. If a share of General Motors stock was worth \$80 at the beginning of the year, fell to \$72 by the end of the year, and paid a dividend of \$4, it will have yielded a return of −5 percent (the dividend yield of 5 percent less the capital loss of 10 percent).

When people invest their savings in stocks, bonds, land, or other assets, they usually hope to earn a return that exceeds the rate of inflation. Thus, by delaying consumption, they can buy more in the future than they can by spending all their income now. Consequently, we often express the return on an asset in *real*—i.e., *inflation-adjusted*—terms. The **real return** on an asset is its simple (or nominal) return *less* the rate of inflation. For example, with an annual inflation rate of 5 percent, our bond, apartment building, and share of GM stock have yielded real returns of 5 percent, 10 percent, and −10 percent, respectively.

• **real return** Simple (or nominal) return on an asset, less the rate of inflation.

EXPECTED VERSUS ACTUAL RETURNS Because most assets are risky, an investor cannot know in advance what returns they will yield over the coming year. For example, our apartment building might have depreciated in value instead of appreciating, and the price of GM stock might have risen instead of fallen. However, we can still compare assets by looking at their expected returns. The **expected return** on an asset is *the expected value of its return*, i.e., the return that it should earn on average. In some years, an asset's **actual return** may be much higher than its expected return and in some years much lower. Over a long period, however, the average return should be close to the expected return.

• **expected return** Return that an asset should earn on average.

• **actual return** Return that an asset earns.

Different assets have different expected returns. Table 5.8, for example, shows that while the expected real return of a U.S. Treasury bill has been less than 1 percent, the expected real return on a group of representative stocks on the New York Stock Exchange has been more than 9 percent.¹² Why would anyone buy a Treasury bill when the expected return on stocks is so much higher? Because the demand for an asset depends not just on its expected return, but also on its *risk*: Although stocks have a higher expected return than Treasury bills, they also carry much more risk. One measure of risk, the standard deviation of the real annual return, is equal to 20.4 percent for common stocks, 8.3 percent for corporate bonds, and only 3.1 percent for U.S. Treasury bills.

The numbers in Table 5.8 suggest that the higher the expected return on an investment, the greater the risk involved. Assuming that one's investments are well diversified, this is indeed the case.¹³ As a result, the risk-averse investor must balance expected return against risk. We examine this trade-off in more detail in the next section.

¹¹The price of a bond often changes during the course of a year. If the bond appreciates (or depreciates) in value during the year, its return will be greater (or less) than 10 percent. In addition, the definition of *return* given above should not be confused with the "internal rate of return," which is sometimes used to compare monetary flows occurring over a period of time. We discuss other return measures in Chapter 15, when we deal with present discounted values.

¹²For some stocks, the expected return is higher, and for some it is lower. Stocks of smaller companies (e.g., some of those traded on the NASDAQ) have higher expected rates of return—and higher return standard deviations.

¹³It is *nondiversifiable* risk that matters. An individual stock may be very risky but still have a low expected return because most of the risk could be diversified away by holding a large number of such stocks. *Nondiversifiable risk*, which arises from the fact that individual stock prices are correlated with the overall stock market, is the risk that remains even if one holds a diversified portfolio of stocks. We discuss this point in detail in the context of the *capital asset pricing model* in Chapter 15.

**TABLE 7.8 INVESTMENTS—RISK AND RETURN (1926–2010)**

	AVERAGE RATE OF RETURN (%)	AVERAGE REAL RATE OF RETURN (%)	RISK (STANDARD DEVIATION)
Common stocks (S&P 500)	11.9	8.7	20.4
Long-term corporate bonds	6.2	3.3	8.3
U.S. Treasury bills	3.7	0.7	3.1

Source: Ibbotson® S&P® 2001 Classic Yearbook: Market results for Stocks, Bonds, Bills, and Inflation 1926–2010.
© 2011 Morningstar.

The Trade-Off Between Risk and Return

Suppose a woman wants to invest her savings in two assets—Treasury bills, which are almost risk free, and a representative group of stocks. She must decide how much to invest in each asset. She might, for instance, invest only in Treasury bills, only in stocks, or in some combination of the two. As we will see, this problem is analogous to the consumer’s problem of allocating a budget between purchases of food and clothing.

Let’s denote the risk-free return on the Treasury bill by R_f . Because the return is risk free, the expected and actual returns are the same. In addition, let the *expected* return from investing in the stock market be R_m and the actual return be r_m . The actual return is risky. At the time of the investment decision, we know the set of possible outcomes and the likelihood of each, but we do not know what particular outcome will occur. The risky asset will have a higher expected return than the risk-free asset ($R_m > R_f$). Otherwise, risk-averse investors would buy only Treasury bills and no stocks would be sold.

THE INVESTMENT PORTFOLIO To determine how much money the investor should put in each asset, let’s set b equal to the fraction of her savings placed in the stock market and $(1 - b)$ the fraction used to purchase Treasury bills. The expected return on her total portfolio, R_p , is a weighted average of the expected return on the two assets:¹⁴

$$R_p = bR_m + (1 - b)R_f \quad (5.1)$$

Suppose, for example, that Treasury bills pay 4 percent ($R_f = .04$), the stock market’s expected return is 12 percent ($R_m = .12$), and $b = 1/2$. Then $R_p = 8$ percent. How risky is this portfolio? One measure of riskiness is the standard deviation of its return. We will denote the *standard deviation* of the risky stock market investment by σ_m . With some algebra, we can show that the *standard deviation of the portfolio*, σ_p (with one risky and one risk-free asset) is the fraction

¹⁴The expected value of the sum of two variables is the sum of the expected values. Therefore

$$R_p = E[br_m] + E[(1 - b)R_f] = bE[r_m] + (1 - b)R_f = bR_m + (1 - b)R_f$$



of the portfolio invested in the risky asset times the standard deviation of that asset.¹⁵

$$\sigma_p = b\sigma_m \quad (5.2)$$

The Investor's Choice Problem

We have still not determined how the investor should choose this fraction b . To do so, we must first show that she faces a risk-return trade-off analogous to a consumer's budget line. To identify this trade-off, note that equation (5.1) for the expected return on the portfolio can be rewritten as

$$R_p = R_f + b(R_m - R_f)$$

Now, from equation (5.2) we see that $b = \sigma_p / \sigma_m$, so that

$$R_p = R_f + \frac{(R_m - R_f)}{\sigma_m} \sigma_p \quad (5.3)$$

RISK AND THE BUDGET LINE This equation is a *budget line* because it describes the trade-off between risk (σ_p) and expected return (R_p). Note that it is the equation for a straight line: Because R_m , R_f , and σ_m are constants, the slope $(R_m - R_f) / \sigma_m$ is a constant, as is the intercept, R_f . The equation says that *the expected return on the portfolio R_p increases as the standard deviation of that return σ_p increases*. We call the slope of this budget line, $(R_m - R_f) / \sigma_m$, the **price of risk**, because it tells us how much extra risk an investor must incur to enjoy a higher expected return.

The budget line is drawn in Figure 5.6. If our investor wants no risk, she can invest all her funds in Treasury bills ($b = 0$) and earn an expected return R_f . To receive a higher expected return, she must incur some risk. For example, she could invest all her funds in stocks ($b = 1$), earning an expected return R_m but incurring a standard deviation σ_m . Or she might invest some fraction of her funds in each type of asset, earning an expected return somewhere between R_f and R_m and facing a standard deviation less than σ_m but greater than zero.

RISK AND INDIFFERENCE CURVES Figure 5.6 also shows the solution to the investor's problem. Three indifference curves are drawn in the figure. Each curve describes combinations of risk and return that leave the investor equally satisfied. The curves are upward-sloping because risk is undesirable. Thus, with a greater amount of risk, it takes a greater expected return to make the investor equally well-off. Curve U_3 yields the greatest amount of satisfaction and U_1 the least amount: For a given amount of risk, the investor earns a higher expected return on U_3 than on U_2 and a higher expected return on U_2 than on U_1 .

¹⁵To see why, we observe from footnote 4 that we can write the variance of the portfolio return as

$$\sigma_p^2 = E[bR_m + (1 - b)R_f - R_p]^2$$

Substituting equation (5.1) for the expected return on the portfolio, R_p , we have

$$\sigma_p^2 = E[bR_m + (1 - b)R_f - bR_m - (1 - b)R_f]^2 = E[b(R_m - R_m)]^2 = b^2\sigma_m^2$$

Because the standard deviation of a random variable is the square root of its variance, $\sigma_p = b\sigma_m$.

In §3.2 we explain how a budget line is determined from an individual's income and the prices of the available goods.

• **Price of risk** Extra risk that an investor must incur to enjoy a higher expected return.

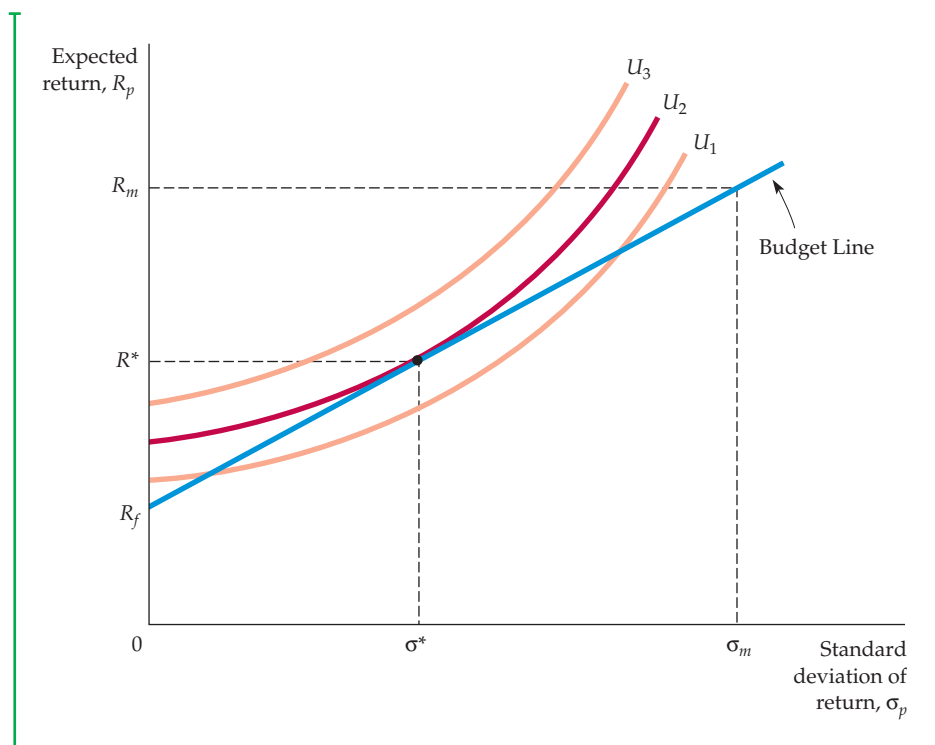


FIGURE 7.6
CHOOSING BETWEEN RISK AND RETURN

An investor is dividing her funds between two assets—Treasury bills, which are risk free, and stocks. The budget line describes the trade-off between the expected return and its riskiness, as measured by the standard deviation of the return. The slope of the budget line is $(R_m - R_f)/\sigma_m$, which is the price of risk. Three indifference curves are drawn, each showing combinations of risk and return that leave an investor equally satisfied. The curves are upward-sloping because a risk-averse investor will require a higher expected return if she is to bear a greater amount of risk. The utility-maximizing investment portfolio is at the point where indifference curve U_2 is tangent to the budget line.

Of the three indifference curves, the investor would prefer to be on U_3 . This position, however, is not feasible, because U_3 does not touch the budget line. Curve U_1 is feasible, but the investor can do better. Like the consumer choosing quantities of food and clothing, our investor does best by choosing a combination of risk and return at the point where an indifference curve (in this case U_2) is tangent to the budget line. At that point, the investor's return has an expected value R^* and a standard deviation σ^* .

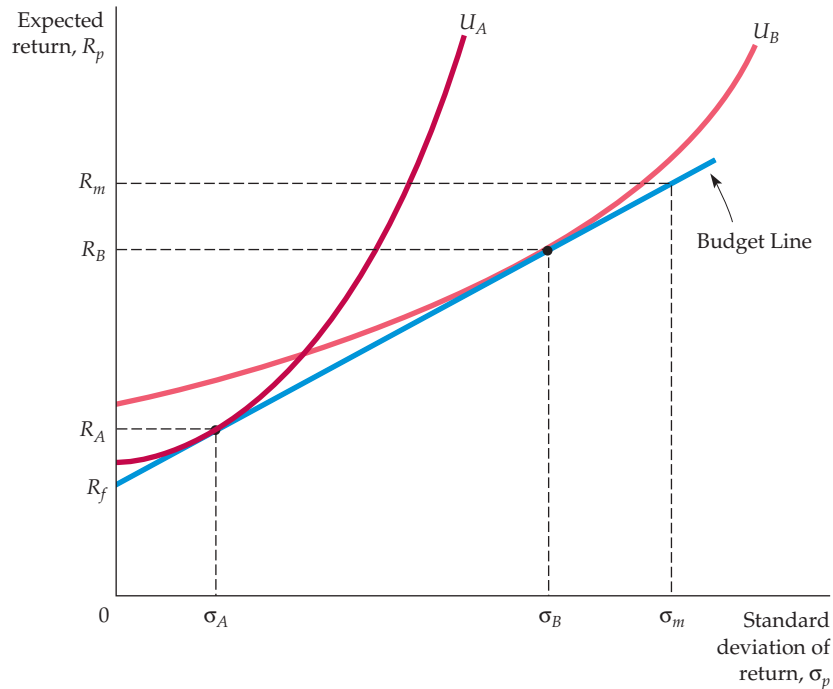
Naturally, people differ in their attitudes toward risk. This fact is illustrated in Figure 5.7, which shows how two different investors choose their portfolios. Investor A is quite risk averse. Because his indifference curve U_A is tangent to the budget line at a point of low risk, he will invest almost all of his funds in Treasury bills and earn an expected return R_A just slightly larger than the risk-free return R_f . Investor B is less risk averse. She will invest most of her funds in stocks, and while the return on her portfolio will have a higher expected value R_B , it will also have a higher standard deviation σ_B .

If Investor B has a sufficiently low level of risk aversion, she might buy stocks on *margin*: that is, she would borrow money from a brokerage firm in order



FIGURE 7.7 THE CHOICES OF TWO DIFFERENT INVESTORS

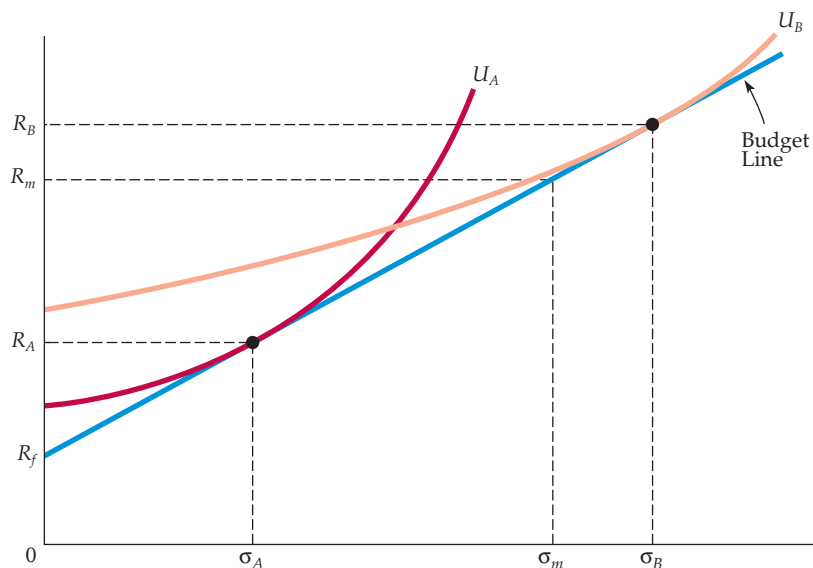
Investor A is highly risk averse. Because his portfolio will consist mostly of the risk-free asset, his expected return R_A will be only slightly greater than the risk-free return. His risk σ_A , however, will be small. Investor B is less risk averse. She will invest a large fraction of her funds in stocks. Although the expected return on her portfolio R_B will be larger, it will also be riskier.



to invest more than she actually owns in the stock market. In effect, a person who buys stocks on margin holds a portfolio with more than 100 percent of the portfolio's value invested in stocks. This situation is illustrated in Figure 5.8, which shows indifference curves for two investors. Investor A, who is relatively risk-averse, invests about half of his funds in stocks. Investor B, however, has an indifference curve that is relatively flat and tangent with the budget line at

FIGURE 7.8 BUYING STOCKS ON MARGIN

Because Investor A is risk averse, his portfolio contains a mixture of stocks and risk-free Treasury bills. Investor B, however, has a very low degree of risk aversion. Her indifference curve, U_B , is tangent to the budget line at a point where the expected return and standard deviation for her portfolio exceed those for the stock market overall. This implies that she would like to invest more than 100 percent of her wealth in the stock market. She does so by buying stocks on margin—i.e., by borrowing from a brokerage firm to help finance her investment.





a point where the expected return on the portfolio exceeds the expected return on the stock market. In order to hold this portfolio, the investor must borrow money because she wants to invest *more* than 100 percent of her wealth in the stock market. Buying stocks on margin in this way is a form of *leverage*: the investor increases her expected return above that for the overall stock market, but at the cost of increased risk.

In Chapters 3 and 4, we simplified the problem of consumer choice by assuming that the consumer had only two goods from which to choose—food and clothing. In the same spirit, we have simplified the investor's choice by limiting it to Treasury bills and stocks. The basic principles, however, would be the same if we had more assets (e.g., corporate bonds, land, and different types of stocks). Every investor faces a trade-off between risk and return.¹⁶ The degree of extra risk that each is willing to bear in order to earn a higher expected return depends on how risk averse he or she is. Less risk-averse investors tend to include a larger fraction of risky assets in their portfolios.

EXAMPLE 7.6 INVESTING IN THE STOCK MARKET

The 1990s witnessed a shift in the investing behavior of Americans. First, many people started investing in the stock market for the first time. In 1989, about 32 percent of families in the United States had part of their wealth invested in the stock market, either directly (by owning indi-



vidual stocks) or indirectly (through mutual funds or pension plans invested in stocks). By 1998, that fraction had risen to 49 percent. In addition, the share of wealth invested in stocks increased from about 26 percent to about 54 percent during the same period.¹⁷ Much of this shift is attributable to younger investors. For those under the age of 35, participation in the stock market increased from about 22 percent in 1989 to about 41 percent in 1998. In most respects, household investing behavior has stabilized after the 1990s shift. The percent of families with investments in the stock market was 51.1% in 2007. However, older Americans have become much more active. By

2007, 40 percent of people over age 75 held stocks, up from 29 percent in 1998.

Why have more people started investing in the stock market? One reason is the advent of online trading, which has made investing much easier. Another reason may be the consider-

able increase in stock prices that occurred during the late 1990s, driven in part by the so-called "dot com euphoria." These increases may have convinced some investors that prices could only continue to rise in the future. As one analyst put it, "The market's relentless seven-year climb, the popularity of mutual funds, the shift by employers to self-directed retirement plans, and the avalanche of do-it-yourself investment publications all have combined to create a nation of financial know-it-alls."¹⁸

Figure 5.9 shows the dividend yield and price/earnings (P/E) ratio for the S&P 500 (an index of stocks of 500 large corporations) over the period

¹⁶As mentioned earlier, what matters is nondiversifiable risk, because investors can eliminate diversifiable risk by holding many different stocks (e.g., via mutual funds). We discuss diversifiable versus nondiversifiable risk in Chapter 15.

¹⁷Data are from the *Federal Reserve Bulletin*, January 2000, and the *Survey of Consumer Finances*, 2011.

¹⁸"We're All Bulls Here: Strong Market Makes Everybody an Expert," *Wall Street Journal*, September 12, 1997.

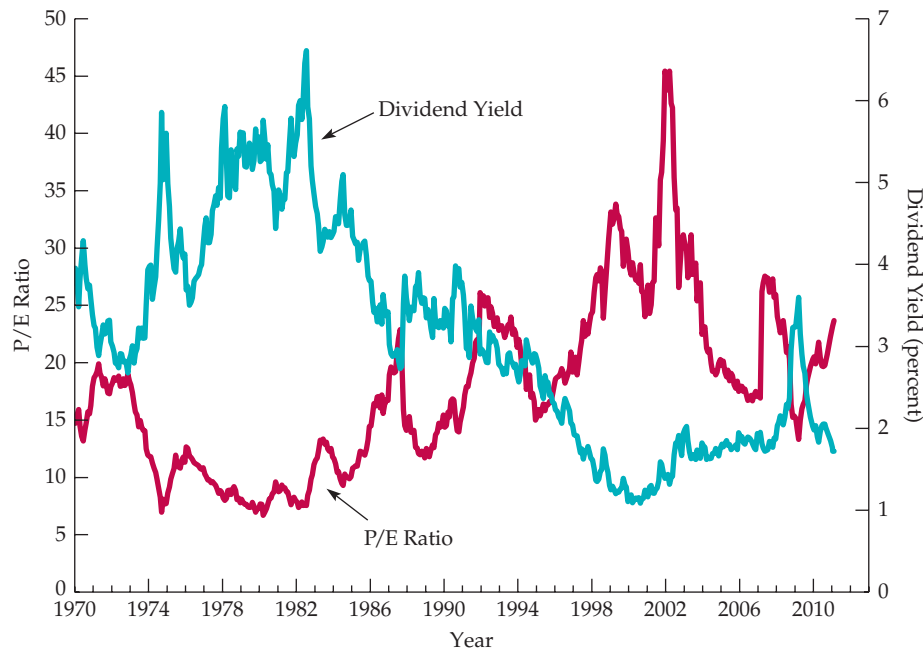


FIGURE 7.9
DIVIDEND YIELD AND P/E RATIO FOR S&P 500

The dividend yield for the S&P 500 (the annual dividend divided by the stock price) has fallen dramatically, while the price/earnings ratio (the stock price divided by the annual earnings-per-share) rose from 1980 to 2002 and then dropped.

1970 to 2011. Observe that the dividend yield (the annual dividend divided by the stock price) fell from about 5 percent in 1980 to below 2 percent by 2000. Meanwhile, however, the price/earnings ratio (the share price divided by annual earnings per share) increased from about 8 in 1980 to over 40 in 2002, before falling to around 20 between 2005 and 2007 and then increasing through 2011. In retrospect, the increase in the P/E ratio could only have occurred if investors believed that corporate profits would continue to grow rapidly in the coming decade. This suggests that in the late 1990s, many investors had a low degree of risk aversion, were quite optimistic about the economy, or both. Alternatively, some economists have argued that the run-up of stock

prices during the 1990s was the result of “herd behavior,” in which investors rushed to get into the market after hearing of the successful experiences of others.¹⁹

The psychological motivations that explain herd behavior can help to explain stock market bubbles. However, they go far beyond the stock market. They also apply to the behavior of consumers and firm managers in a wide variety of settings. Such behavior cannot always be captured by the simplified assumptions that we have made up to this point about consumer choice. In the next section, we will discuss these aspects of behavior in detail, and we will see how the traditional models of Chapters 3 and 4 can be expanded to help us understand this behavior.

¹⁹See, for example, Robert Shiller, *Irrational Exuberance*, Princeton University Press, 2000.



7.5 Bubbles

During 1995 to 2000, the stock prices of many Internet companies rose sharply. What was behind these sharp price increases? One could argue—as many stock analysts, investment advisors, and ordinary investors did at the time—that these price increases were justified by fundamentals. Many people thought that the Internet’s potential was virtually unbounded, particularly as high-speed Internet access became more widely available. After all, more and more goods and services were being bought online through companies such as Amazon.com, Craigslist.org, Ticketmaster.com, Fandango.com, and a host of others. In addition, more and more people began to read the news online rather than buying physical newspapers and magazines, and more and more information became available online through sources like Google, Bing, Wikipedia, and WebMD. And as a result, companies began to shift more and more of their advertising from newspapers and television to the Internet.

Yes, the Internet has certainly changed the way most of us live. (In fact, some of you may be reading the electronic version of this book, which you downloaded from the Pearson website and hopefully paid for!) But does that mean that any company with a name that ends in “.com” is sure to make high profits in the future? Probably not. And yet many investors (perhaps “speculators” is a better word) bought the stocks of Internet companies at very high prices, prices that were increasingly difficult to justify based on fundamentals, i.e., based on rational projections of future profitability. The result was the Internet **bubble**, an increase in the prices of Internet stocks based not on the fundamentals of business profitability, but instead on the belief that the prices of those stocks would keep going up. The bubble burst when people started to realize that the profitability of these companies was far from a sure thing, and that prices that go up can also come down.

Bubbles are often the result of irrational behavior. People stop thinking straight. They buy something because the price has been going up, and they believe (perhaps encouraged by their friends) that the price will keep going up, so that making a profit is a sure thing. If you ask these people whether the price might at some point drop, they typically will answer “Yes, but I will sell before the price drops.” And if you push them further by asking how they will know when the price is about to drop, the answer might be “I’ll just know.” But, of course, most of the time they won’t know; they will sell after the price has dropped, and they will lose at least part of their investment. (There might be a silver lining—perhaps they will learn some economics from the experience.)

Bubbles are often harmless in the sense that while people lose money, there is no lasting damage to the overall economy. But that is not always the case. The United States experienced a prolonged housing price bubble that burst in 2008, causing financial losses to large banks that had sold mortgages to home buyers who could not afford to make their monthly payments (but thought housing prices would keep rising). Some of these banks were given large government bailouts to keep them from going bankrupt, but many homeowners were less fortunate, and facing foreclosure, they lost their homes. By the end of 2008, the United States was in its worst recession since the Great Depression of the 1930s. The housing price bubble, far from harmless, was partly to blame for this.

• **bubble** An increase in the price of a good based not on the fundamentals of demand or value, but instead on a belief that the price will keep going up.

Recall from Section 4.3 that *speculative demand* is driven not by the direct benefits one obtains from owning or consuming a good but instead is driven by an expectation that the price of the good will increase.



EXAMPLE 7.7 THE HOUSING PRICE BUBBLE (I)

Starting around 1998, U.S. housing prices began rising sharply. Figure 5.10 shows the S&P/Case-Shiller housing price index at the national level.²⁰ From 1987 (when the Index was first published) to 1998, the index rose around 3 percent per year in nominal terms. (In real terms, i.e., net of inflation, the index dropped about 0.5 percent per year.) This was a normal rate of price increase, roughly commensurate with population and income growth and with inflation. But then prices started rising much more rapidly, with the index increasing about 10 percent per year until it reached its peak of 190 in 2006. During that 8-year period from 1998 to 2006,



many people bought into the myth that housing was a sure-fire investment, and that prices could only keep going up. Many banks also bought into this myth and offered mortgages to people with incomes well below what it would take to make the monthly interest and principal payments over

the long term. The demand for housing increased sharply, with some people buying four or five houses under the assumption that they could “flip” them in a year and make a quick profit. This speculative demand served to push prices up further.

However, in 2006 something funny happened. Prices stopped going up. In fact, during 2006, prices

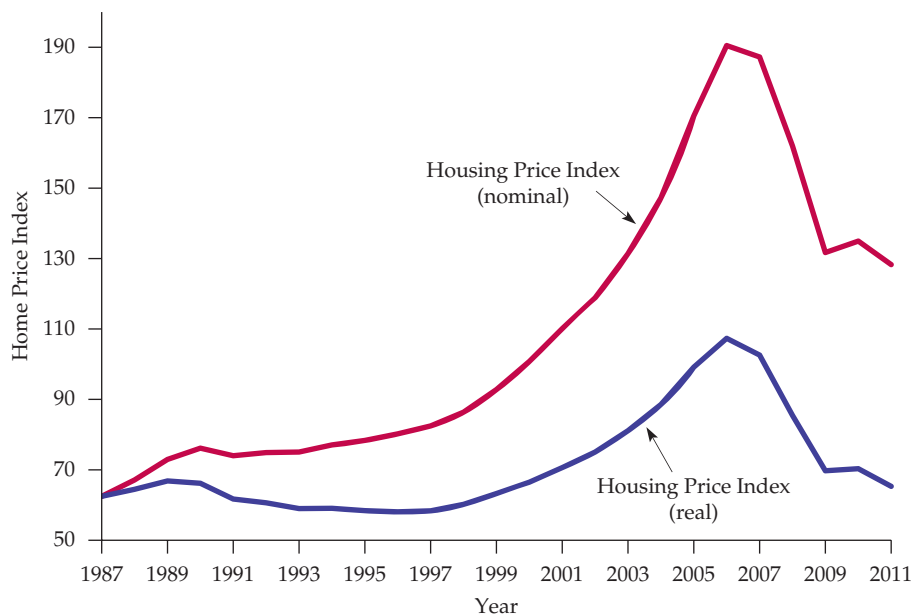


FIGURE 7.10
S&P/CASE-SHILLER HOUSING PRICE INDEX

The Index shows the average home price in the United States at the national level. Note the increase in the index from 1998 to 2007, and then the sharp decline.

²⁰The S&P/Case-Shiller index measures the change in housing prices by tracking repeat sales of single family homes in 20 cities across the United States. By comparing a home's original sale price with its price in subsequent sales, the index is able to control for other variables (i.e., size, location, style) that might also lead to rising home prices.



actually fell slightly (about 2 percent in nominal terms). Then, in 2007 prices started falling rapidly, and by 2008 it had become clear that the great housing boom was just a bubble, and the bubble had burst. From its peak in early 2006 through 2011, housing prices fell by over 33 percent in nominal terms. (In real terms they fell by nearly 40 percent.) And this drop is an average for the United States as a whole. In some states, such as Florida, Arizona, and Nevada, the bubble was far worse, with prices dropping by over 50 percent.

The United States was not the only country to experience a housing price bubble. More or less the same thing happened in Europe. In Ireland, for

example, a booming economy and increasing foreign investment—along with widespread speculation—pushed housing prices up 305% between 1995 and 2007 (641% between 1987 and 2007—both in nominal terms). After over a decade of above average growth, Ireland's bubble burst. By 2010, housing prices had fallen over 28% from their 2007 peak. Spain and other European countries suffered similar fates, contributing to a worldwide debt crisis. Other apparent bubbles have yet to deflate. Many Chinese cities, including Shanghai and Beijing, have seen rapidly rising housing and land prices, with some apartments reportedly doubling in value in mere months.²¹

Informational Cascades

Suppose you are considering investing in the stock of Ajax Corp., which is trading at \$20 per share. Ajax is a biotech company that is working on a radically new approach to the treatment of chronic boredom (a disease that often afflicts students of economics). You find it difficult to evaluate the company's prospects, but \$20 seems like a reasonable price. But now you see the price is increasing—to \$21, \$22, then a jump to \$25 per share. In fact, some friends of yours have just bought in at \$25. Now the price reaches \$30. Other investors must know something. Perhaps they consulted biochemists who can better evaluate the company's prospects. So you decide to buy the stock at \$30. You believe that positive information drove the actions of other investors, and you acted accordingly.

Was buying the stock of Ajax at \$30 a rational decision, or were you simply buying into a bubble? It might indeed be rational. After all, it is reasonable to expect that other investors tried to value the company as best they could and that their analyses might have been more thorough or better informed than yours. Thus the actions of other investors could well be informative and lead you to rationally adjust your own valuation of the company.

Note that in this example, your investment decisions are based not on fundamental information that you have obtained (e.g., regarding the likelihood that Ajax's R&D will be successful), but rather on the investment decisions of others. And note that you are implicitly assuming that: (i) these investment decisions of others are based on fundamental information that *they* have obtained; or (ii) these investment decisions of others are based on the investment decisions of others still, which are based on fundamental information that *they* have obtained; or (iii) these investment decisions of others are based on the investment decisions of others still, which in turn are based on the investment decisions of still more others, which are based on fundamental information that *they* obtained; or ... etc., etc. You get the idea. Maybe the "others" at the end of the chain based their investment decisions on weak information that was no more informative than the information you started with when you began thinking about Ajax. In other

²¹Fearing a sudden collapse, the Chinese government took steps to curtail skyrocketing housing prices, tightening lending requirements and requiring purchasers to put more money down. See <http://www.businessinsider.com/the-chinese-real-estate-bubble-is-the-most-obvious-bubble-ever-2010-1#prices-are-way-out-of-whack-compared-to-global-standards-3>.



EXAMPLE 7.8 THE HOUSING PRICE BUBBLE (II)

Informational cascades may help to explain the housing bubbles that occurred in the U.S. and other countries. For example, from 1999 to 2006, home prices in Miami nearly tripled. Would it have been completely irrational to buy real estate in Miami in 2006? In the years prior to 2006, some analysts projected large increases in the demand for housing in Miami and other parts of Florida, based in part on a growing number of aging retirees that want to move to someplace warm, and in part on an influx of immigrants with family or other roots in Miami. If other investors acted on the belief that these analysts had done their homework, investing might have been rational.



Informational cascades may also help explain the housing bubbles that took place in other parts of the U.S., notably Arizona, Nevada, and California. (See Figure 5.11.) There, too, some analysts had projected large increases in demand. On the other hand, few analysts projected large demand increases in cities like Cleveland (not exactly a retirement paradise), and indeed such cities experienced little in the way of a bubble.

Was it rational to buy real estate in Miami in 2006? Rational or not, investors should have known that considerable risk was involved in buying real estate there (or elsewhere in Florida, Arizona, Nevada, and California). Looking back, we now know that many of these investors lost their shirts (not to mention their homes).

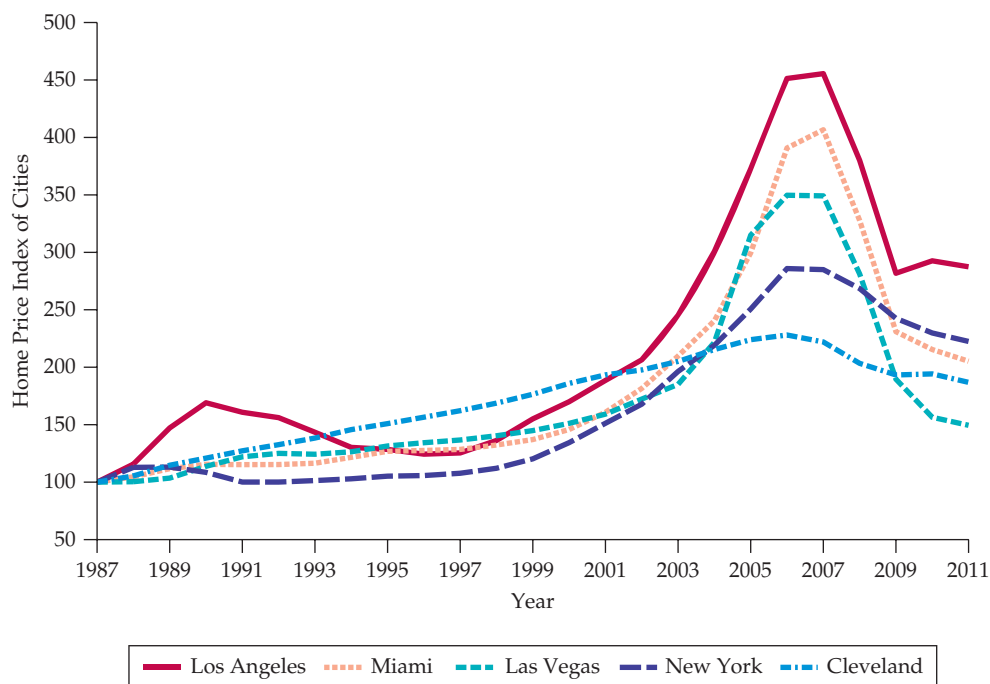


FIGURE 7.11
S&P/CASE-SHILLER HOUSING PRICE INDEX FOR FIVE CITIES

The Index shows the average home price for each of five cities (in nominal terms). For some cities, the housing bubble was much worse than for others. Los Angeles, Miami, and Las Vegas experienced some of the sharpest increases in home prices, and then starting in 2007, prices plummeted. Cleveland, on the other hand, largely avoided the bubble, with home prices increasing, and then falling, only moderately.



words, your own investment decisions might be the result of an **informational cascade**—actions based on actions based on actions . . . , etc., driven by very limited fundamental information.

The bubble that results from an informational cascade can in fact be rational in the sense that there is a basis for believing that investing in the bubble will yield a positive return. The reason is that if investors early in the chain indeed obtained positive information and based their decisions on that information, *the expected gain to an investor down the chain will be positive*.²² However, the risk involved will be considerable, and it is likely that at least some investors will underestimate that risk.

• **Informational cascade**

An assessment (e.g., of an investment opportunity) based in part on the actions of others, which in turn were based on the actions of others.

7.6 Behavioral Economics

Recall that the basic theory of consumer demand is based on three assumptions: (1) consumers have clear preferences for some goods over others; (2) consumers face budget constraints; and (3) given their preferences, limited incomes, and the prices of different goods, consumers choose to buy combinations of goods that maximize their satisfaction (or utility). These assumptions, however, are not always realistic: Preferences are not always clear or might vary depending on the context in which choices are made, and consumer choices are not always utility-maximizing.

Perhaps our understanding of consumer demand (as well as the decisions of firms) would be improved if we incorporated more realistic and detailed assumptions regarding human behavior. This has been the objective of the newly flourishing field of *behavioral economics*, which has broadened and enriched the study of microeconomics.²³ We introduce this topic by highlighting some examples of consumer behavior that cannot be easily explained with the basic utility-maximizing assumptions that we have relied on so far:

- There has just been a big snowstorm, so you stop at the hardware store to buy a snow shovel. You had expected to pay \$20 for the shovel—the price that the store normally charges. However, you find that the store has suddenly raised the price to \$40. Although you would expect a price increase because of the storm, you feel that a doubling of the price is unfair and that the store is trying to take advantage of you. Out of spite, you do not buy the shovel.²⁴
- Tired of being snowed in at home you decide to take a vacation in the country. On the way, you stop at a highway restaurant for lunch. Even though you are unlikely to return to that restaurant, you believe that it is fair and

²²For a reasonably simple example that makes this point (and an interesting discussion), see S. Bikhchandani, D. Hirschleifer, and I. Welch, “Learning from the Behavior of Others: Conformity, Fads, and Informational Cascades,” 12 *Journal of Economic Perspectives*, (Summer 1998): 151–170.

²³For more detailed discussion of the material presented in this section, see Stefano DellaVigna, “Psychology and Economics: Evidence from the Field,” *Journal of Economic Literature* 47(2), 2009: 315–372; Colin Camerer and George Loewenstein, “Behavioral Economics: Past, Present, Future,” in Colin Camerer, George Loewenstein, and Matthew Rabin (eds.), *Advances in Behavioral Economics*, Princeton University Press, 2003.

²⁴This example is based on Daniel Kahneman, Jack Knetsch, and Richard Thaler, “Fairness as a Constraint on Profit Seeking: Entitlements in the Market,” *American Economic Review* 76 (September 1986): 728–741.



appropriate to leave a 15-percent tip in appreciation of the good service that you received.

- You buy this textbook from an Internet bookseller because the price is lower than the price at your local bookstore. However, you ignore the shipping cost when comparing prices.

Each of these examples illustrates plausible behavior that cannot be explained by a model based solely on the basic assumptions described in Chapters 3 and 4. Instead, we need to draw on insights from psychology and sociology to augment our basic assumptions about consumer behavior. These insights will enable us to account for more complex consumer preferences, for the use of simple rules in decision-making, and for the difficulty that people often have in understanding the laws of probability.

Adjustments to the standard model of consumer preferences and demand can be grouped into three categories: A tendency to value goods and services in part based on the setting one is in, a concern about the fairness of an economic transaction, and the use of simple rules of thumb as a way to cut through complex economic decisions. We examine each of these in turn.

Reference Points and Consumer Preferences

The standard model of consumer behavior assumes that consumers place unique values on the goods and services they purchase. However, psychologists and market research studies have found that perceived value depends in part on the setting in which the purchasing decision occurs. That setting creates a **reference point** on which preferences might be at least partly based.

The reference point—the point from which the individual makes the consumption decision—can strongly affect that decision. Consider, for example, apartment prices in Pittsburgh and San Francisco. In Pittsburgh, the median monthly rent in 2006 for a two-bedroom apartment was about \$650, while in San Francisco the rent for a similar apartment was \$2,125. For someone accustomed to San Francisco housing prices, Pittsburgh might seem like a bargain. On the other hand, someone moving from Pittsburgh to San Francisco might feel “gouged”—thinking it unfair for housing to cost that much.²⁵ In this example, the reference point is clearly different for long-time residents of Pittsburgh and San Francisco.

Reference points can develop for many reasons: our past consumption of a good, our experience in a market, our expectation about how prices should behave, and even the context in which we consume a good. Reference points can strongly affect the way people approach economic decisions. Below we describe several different examples of reference points and the way they affect consumer behavior.

ENDOWMENT EFFECT A well-known example of a reference point is the **endowment effect**—the fact that individuals tend to value an item more when they happen to own it than when they do not. One way to think about this effect is to consider the gap between the price that a person is willing to pay for a good and the price at which she is willing to sell the same good to someone else. Our

• **reference point** The point from which an individual makes a consumption decision.

• **endowment effect** Tendency of individuals to value an item more when they own it than when they do not.

²⁵This example is based on Uri Simonsohn and George Loewenstein, “Mistake #37: The Effects of Previously Encountered Prices on Current Housing Demand,” *The Economic Journal* 116 (January 2006): 175–199.



basic theory of consumer behavior says that this price should be the same, but many experiments suggest that is not what happens in practice.²⁶

In one classroom experiment, half of the students chosen at random were given a free coffee mug with a market value of \$5; the other half got nothing.²⁷ Students with the mug were asked the price at which they would sell it back to the professor; the second group was asked the minimum amount of money that they would accept in lieu of a mug. The decision faced by both groups is similar but their reference points are different. For the first group, whose reference point was possession of a mug, the average selling price was \$7. For the second group, which did not have a mug, the average amount desired in lieu of a mug was \$3.50. This gap in prices shows that giving up the mug was perceived to be a greater “loss” to those who had one than the “gain” from obtaining a mug for those without one. This is an endowment effect—the mug was worth more to those people who already owned it.

LOSS AVERSION The coffee mug experiment described above is also an example of **loss aversion**—the tendency of individuals to prefer avoiding losses over acquiring gains. The students who owned the mug and believed that its market value was indeed \$5 were averse to selling it for less than \$5 because doing so would have created a perceived loss. The fact that they had been given the mug for free, and thus would still have had an overall gain, didn’t matter as much.

As another example of loss aversion, people are sometimes hesitant to sell stocks at a loss, even if they could invest the proceeds in other stocks that they think are better investments. Why? Because the original price paid for the stock—which turned out to be too high given the realities of the market—acts as a reference point, and people are averse to losses. (A \$1000 loss on an investment seems to “hurt” more than the perceived benefit from a \$1000 gain.) While there are a variety of circumstances in which endowment effects arise, we now know that these effects tend to disappear as consumers gain relevant experience. We would not expect to see stockbrokers or other investment professionals exhibit the loss aversion described above.²⁸

FRAMING Preferences are also influenced by **framing**, which is another manifestation of reference points. Framing is a tendency to rely on the context in which a choice is described when making a decision. How choices are framed—the names they are given, the context in which they are described, and their appearance—can affect the choices that individuals make. Are you more likely to buy a skin cream whose package claims that it will “slow the aging process” or one that is described as “making you feel young again.” These products might be essentially identical except for their packaging. Yet, in the real world where information is sometimes limited and perspective matters, many individuals would prefer to buy the product that emphasizes youth.

• **loss aversion** Tendency for individuals to prefer avoiding losses over acquiring gains.

• **framing** Tendency to rely on the context in which a choice is described when making a decision.

²⁶Experimental work such as this has been important to the development of behavioral economics. It is for this reason that the 2002 Nobel Prize in economics was shared by Vernon Smith, who did much of the pioneering work in the use of experiments to test economic theories.

²⁷Daniel Kahneman, Jack L. Knetsch, and Richard H. Thaler, “Experimental Tests of the Endowment Effect and the Coase Theorem,” *Journal of Political Economy* 98, (December 1990): 1925–48.

²⁸John A. List, “Does Market Experience Eliminate Market Anomalies?” *Quarterly Journal of Economics* 118 (January 2003): 41–71.



EXAMPLE 7.9 SELLING A HOUSE

Homeowners sometimes sell their homes because they have to relocate for a new job, because they want to be closer to (or farther from) the city in which they work, or because they want to move to a bigger or smaller house. So they put their home on the market. But at what price? The owners can usually get a good idea of what the house will sell for by looking at the selling prices of comparable houses, or by talking with a realtor. Often, however, the owners will set an asking price that is well above any realistic expectation of what the house can actually sell for. As a result, the house may stay on the market for many months before the owners grudgingly lower the price. During that time the owners have to continue to maintain the house and pay for taxes, utilities, and insurance. This seems irrational. Why not set an asking price closer to what the market will bear?

The *endowment effect* is at work here. The homeowners view their house as special; their

ownership has given them what they think is a special appreciation of its value—a value that may go beyond any price that the market will bear.

If housing prices have been falling, *loss aversion* could also be at work. As we saw in Examples 5.7 and 5.8, U.S. and European housing prices started falling around 2008, as the housing bubble deflated. As a result, some homeowners were affected by loss aversion when deciding on an asking price, especially if they bought their home at a time near the peak of the bubble. Selling the house turns a paper loss, which may not seem real, into a loss that is real. Averting that reality may serve to explain the reluctance of home owners to take that final step of selling their home. It is not surprising, therefore, to find that houses tend to stay on the market longer during economic downturns than in upturns.

Fairness

People sometimes do things because they think it is appropriate or *fair* to do so, even though there is no financial or other material benefit. Examples include charitable giving, volunteering time, or tipping in a restaurant. Fairness likewise affected consumer behavior in our example of buying a snow shovel.

At first glance, our basic consumer theory does not appear to account for fairness. However, we can often modify our models of demand to account for the effects of fairness on consumer behavior. To see how, let's return to our original snow shovel example. In that example, the market price of shovels was \$20, but right after a snowstorm (which caused a shift in the demand curve), stores raised their price to \$40. Some consumers, however, felt they were being unfairly gouged, and refused to buy a shovel.

This is illustrated in Figure 5.12. Demand curve D_1 applies during normal weather. Stores have been charging \$20 for a shovel, and sell a total quantity of Q_1 shovels per month (because many consumers buy shovels in anticipation of snow). In fact some people would have been willing to pay much more for a shovel (the upper part of the demand curve), but they don't have to because the market price is \$20. Then the snowstorm hits, and the demand curve shifts to the right. Had the price remained \$20, the quantity demanded would have increased to Q_2 . But note that the new demand curve (D_2) does not extend up as far as the old one. Many consumers might feel that an increase in price to, say, \$25 is fair, but an increase much above that would be unfair gouging. Thus the new demand curve becomes very elastic at prices above \$25, and no shovels can be sold at a price much above \$30.

Note how fairness comes in to play here. In normal weather, some consumers would have been willing to pay \$30 or even \$40 for a shovel. But they know that

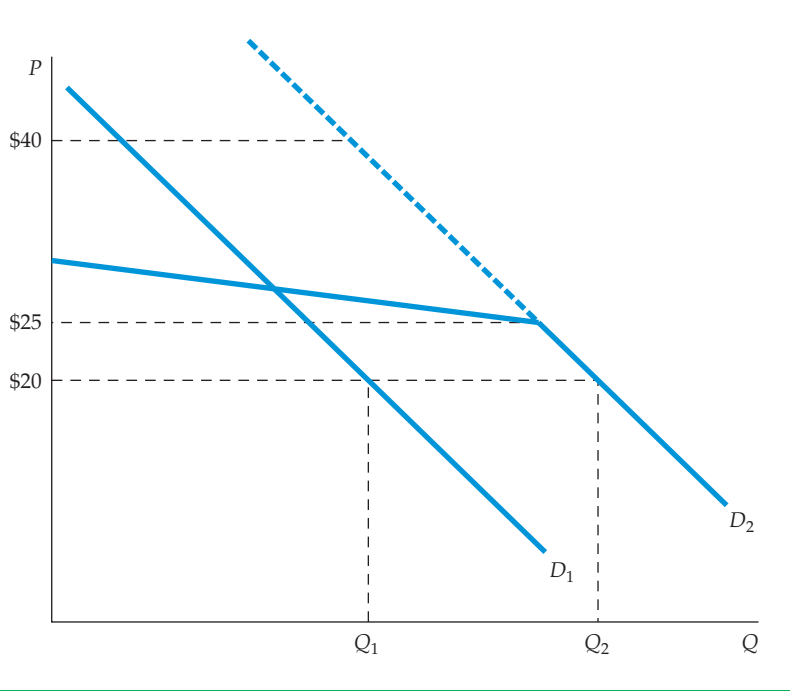


FIGURE 7.12
DEMAND FOR SNOW SHOVELS

Demand curve D_1 applies during normal weather. Stores have been charging \$20 and sell Q_1 shovels per month. When a snowstorm hits, the demand curve shifts to the right. Had the price remained \$20, the quantity demanded would have increased to Q_2 . But the new demand curve (D_2) does not extend up as far as the old one. Consumers view an increase in price to, say, \$25 as fair, but an increase much above that as unfair gouging. The new demand curve is very elastic at prices above \$25, and no shovels can be sold at a price much above \$30.

the price has always been \$20, and they feel that a sharp increase in price after a snowstorm is unfair gouging and refuse to buy. Note also how we can modify standard demand curves to account for consumer attitudes towards fairness.

Another example of fairness arises in the *ultimatum game*. Imagine that, under the following rules, you are offered a chance to divide 100 one-dollar bills with a stranger whom you will never meet again: You first propose a division of the money between you and the stranger. The stranger will respond by either accepting or rejecting your proposal. If he accepts, you each get the share that you proposed. If he rejects, you both get nothing. What should you do?

Because more money means more utility, our basic theory provides a clear answer to this question. You should propose that you get \$99 while the other person gets only \$1. Moreover, the responder should be happy to accept this proposal, because \$1 is more than he had before and more than he would get if he rejected your offer (in both cases zero). This is a beneficial deal for both of you.

Yet most people facing this choice hesitate to make such an offer because they think it unfair, and many “strangers” would reject the offer. Why? The stranger might believe that because you both received the windfall opportunity to divide \$100, a simple and fair division would be 50/50 or something close to that. Maybe the stranger will turn down the \$1 offer to teach you that greediness is not appropriate behavior. Indeed, if you believe that the stranger will feel this way, it will be rational for you to offer a greater amount. In fact, when this game is played experimentally, typical sharing proposals range between 67/33 and 50/50, and such offers are normally accepted.

The ultimatum game shows how fairness can affect economic decisions. Not surprisingly, fairness concerns can also affect negotiations between firms and their workers. A firm may offer a higher wage to employees because the managers believe that workers deserve a comfortable standard of living or because they want to foster a pleasant working environment. Moreover, workers who do



not get a wage that they feel is fair may not put much effort into their work.²⁹ (In Section 17.6, we will see that paying workers higher-than-market wages can also be explained by the “efficiency wage theory” of labor markets, in which fairness concerns do not apply.) Fairness also affects the ways in which firms set prices and can explain why firms can more easily raise prices in response to higher costs than to increases in demand.³⁰

Fortunately, fairness concerns can be taken into account in the basic model of consumer behavior. If individuals moving to San Francisco believe that high apartment rents are unfair, their maximum willingness to pay for rental housing will be reduced. If a sufficient number of individuals feel this way, the resulting reduction in demand will lead to lower rental prices. Similarly, if enough workers do not feel that their wages are fair, there will be a reduction in the supply of labor, and wage rates will increase.

Rules of Thumb and Biases in Decision Making

Many economic (and everyday) decisions can be quite complex, especially if they involve choices about matters in which we have little experience. In such cases, people often resort to rule of thumb or other mental shortcuts to help them make decisions. In the tipping example, you took a mental shortcut when you decided to offer a 15-percent tip. The use of such rules of thumb, however, can introduce a bias into our economic decision making—something that our basic model does not allow.³¹

ANCHORING The mental rules that we use in making decisions frequently depend on both the context in which the decisions are made and the information available. For example, imagine that you just received a solicitation from a new local charity to make a donation. Rather than asking for a gift of any amount, the charity asks you to choose: \$20, \$50, \$100, \$250, or “other.” The purpose of these suggestions is to induce you to anchor your final donation. **Anchoring** refers to the impact that a suggested (perhaps unrelated) piece of information may have on your final decision. Rather than trying to decide precisely how much to donate—say \$44.52—and not wanting to appear miserly, one might simply write a check for the next higher category—\$50. Another individual wishing to make only a token donation of \$10 might choose the lowest stated amount, \$20. In both cases, anchoring can bias individual choices toward larger donations. Similarly, it’s no coincidence so many price tags end with the digits 95 or 99. Marketers understand that consumers tend to overemphasize the first digit of prices, and also to think in terms of price categories like “under \$20” or “over \$20.” Thus to the consumer, who may not be thinking too carefully, \$19.95 seems much cheaper than \$20.01.

• **anchoring** Tendency to rely heavily on one prior (suggested) piece of information when making a decision.

RULES OF THUMB A common way to economize on the effort involved in making decisions is to ignore seemingly unimportant pieces of information.

²⁹For a general discussion of behavioral economics and the theory of wages and employment, see George Akerlof, “Behavioral Macroeconomics and Macroeconomic Behavior,” *American Economic Review* 92 (June 2002): 411–33.

³⁰See, for example, Julio J. Rotemberg, “Fair Pricing,” NBER Working Paper No. W10915, 2004.

³¹For an introduction to this topic see Amos Tversky and Daniel Kahneman, “Judgment under Uncertainty: Heuristics and Biases,” *Science* 185 (September 1974): 1124–31.



For example, goods purchased over the Internet often involve shipping costs. Although small, these costs should be included as part of the good's final price when making a consumption decision. However, a recent study has shown that shipping costs are typically ignored by many consumers when deciding to buy things online. Their decisions are biased because they view the price of goods to be lower than they really are.³²

Whereas depending on rules of thumb can introduce biases in decision making, it is important to understand that they do serve a useful purpose. Frequently, rules of thumb help to save time and effort and result in only small biases. Thus, they should not be dismissed outright.

Consumers often face uncertainty when making decisions, and lack the understanding of probability to make those decisions optimally. (Consider the difficulty involved, for example, in calculating expected utility.) Consumers will often use rules of thumb when making decisions, but sometimes those rules of thumb can lead to strong biases.

THE LAW OF SMALL NUMBERS People are sometimes prone to a bias called the **law of small numbers**: They tend to overstate the probability that certain events will occur when faced with relatively little information from recent memory. For example, many people tend to overstate the likelihood that they or someone they know will die in a plane crash or win the lottery. Recall the roulette player who bets on black after seeing red come up three times in a row: He has ignored the laws of probability.

• **law of small numbers** Tendency to overstate the probability that a certain event will occur when faced with relatively little information.

Research has shown that investors in the stock market are often subject to a small-numbers bias, believing that high returns over the past few years are likely to be followed by more high returns over the next few years—thereby contributing to the kind of “herd behavior” that we discussed in the previous section. In this case, investors assess the likely payoff from investing by observing the market over a short period of time. In fact, one would have to study stock market returns for many decades in order to estimate accurately the expected return on equity investments. Similarly when people assess the likelihood that housing prices will rise based on several years of data, the resulting misperceptions can result in housing price bubbles.³³

Although individuals may have some understanding of true probabilities (as when flipping a coin), complications arise when probabilities are unknown. For instance, few people have an idea about the probability that they or a friend will be in a car or airplane accident. In such cases, we form subjective probability assessments about such events. Our estimation of subjective probabilities may be close to true probabilities, but often they are not.

Forming subjective probabilities is not always an easy task and people are generally prone to several biases in the process. For instance, when evaluating the likelihood of an event, the context in which the evaluation is made can be very important. If a tragedy such as a plane crash has occurred recently, many people will tend to overestimate the probability of it happening to them. Likewise, when a probability for a particular event is very, very small, many people simply ignore that possibility in their decision making.

³²Tankim Hossain and John Morgan, “... Plus Shipping and Handling: Revenue (Non) Equivalence in Field Experiments on eBay,” *Advances in Economic Analysis & Policy* 6: 2 (2006).

³³See Charles Himmelberg, Christopher Mayer, and Todd Sinai, “Assessing High House Prices: Bubbles, Fundamentals and Misperceptions,” *Journal of Economic Perspectives* 19 (Fall 2005): 67–92.



Summing Up

Where does this leave us? Should we dispense with the traditional consumer theory discussed in Chapters 3 and 4? Not at all. In fact, the basic theory that we learned up to now works quite well in many situations. It helps us to understand and evaluate the characteristics of consumer demand and to predict the impact on demand of changes in prices or incomes. Although it does not explain all consumer decisions, it sheds light on many of them. The developing field of behavioral economics tries to explain and to elaborate on those situations that are not well explained by the basic consumer model.

If you continue to study economics, you will notice many cases in which economic models are not a perfect reflection of reality. Economists have to carefully decide, on a case-by-case basis, what features of the real world to include and what simplifying assumptions to make so that models are neither too complicated to study nor too simple to be useful.

EXAMPLE 7.10 NEW YORK CITY TAXICAB DRIVERS

Most cab drivers rent their taxicabs for a fixed daily fee from a company that owns a fleet of cars. They can then choose to drive the cab as little or as much as they want during a 12-hour period. As with many services, business is highly variable from day to day, depending on the weather, subway breakdowns, holidays, and so

on. How do cabdrivers respond to these variations, many of which are largely unpredictable?

In many cities, taxicab rates are fixed by regulation and do not change from day to day. However, on busy days drivers can earn a higher income because they do not have to spend as much time searching for riders. Traditional economic theory would predict that drivers will work longer hours on busy days than on slow days; an extra hour on a busy day might bring in \$20, whereas an extra hour on a slow day might yield only \$10. Does traditional theory explain the actual behavior of taxicab drivers?

An interesting study analyzed actual taxicab trip records obtained from the New York Taxi and Limousine Commission for the spring of 1994.³⁴ The daily fee to rent a taxi was then \$76, and gasoline cost about \$15 per day. Surprisingly, the researchers



found that most drivers drive *more* hours on slow days and *fewer* hours on busy days. In other words, there is a *negative relationship* between the effective hourly wage and the number of hours worked each day; the higher the wage, the sooner the cabdrivers quit for the day. Behavioral economics can

explain this result. Suppose that most taxicab drivers have an income target for each day. That target effectively serves as a reference point. Daily income targeting makes sense from a behavioral perspective. An income target provides a simple decision rule for drivers because they need only keep a record of their fares for the day. A daily target also helps drivers with potential self-control problems; without a target, a driver may choose to quit earlier on many days just to avoid the hassles of the job. The target in the 1994 study appeared to be about \$150 per day.

Still other studies challenge this “behavioral” explanation of behavior. A different study, also of New York City cab drivers who rented their taxis, concluded that the traditional economic model does indeed offer important insights into drivers’

³⁴Colin Camerer, Linda Babcock, George Loewenstein, and Richard Thaler, “Labor Supply of New York City Cabdrivers: One Day at a Time,” *Quarterly Journal of Economics* (May 1997): 404–41. See also, Henry S. Farber, “Reference-Dependent Preferences and Labor Supply: The Case of New York City Taxi Drivers,” *American Economic Review* 98 (2008): 1069–82.



behavior.³⁵ The study concluded that daily income had only a small effect on a driver's decision as to when to quit for the day. Rather, the decision to stop appears to be based on the cumulative number of hours already worked that day and not on hitting a specific income target.

What may soon become known as "the great taxicab driver debate" did not end here. A recent study sought to explain these two seemingly

contradictory results. Reanalyzing the same taxicab trip records, the authors found that the traditional economic model goes a long way in explaining most workday decisions of taxicab drivers, but that a behavioral model that accounts for reference points and targeted goals (for income and hours) can do even better.³⁶ If you are interested in learning more about the taxicab industry, you can look ahead to the examples in Chapters 8, 9, and 15.

SUMMARY

1. Consumers and managers frequently make decisions in which there is uncertainty about the future. This uncertainty is characterized by the term *risk*, which applies when each of the possible outcomes and its probability of occurrence is known.
2. Consumers and investors are concerned about the expected value and the variability of uncertain outcomes. The expected value is a measure of the central tendency of the values of risky outcomes. Variability is frequently measured by the standard deviation of outcomes, which is the square root of the probability-weighted average of the squares of the deviation from the expected value of each possible outcome.
3. Facing uncertain choices, consumers maximize their expected utility—an average of the utility associated with each outcome—with the associated probabilities serving as weights.
4. A person who would prefer a certain return of a given amount to a risky investment with the same expected return is risk averse. The maximum amount of money that a risk-averse person would pay to avoid taking a risk is called the *risk premium*. A person who is indifferent between a risky investment and the certain receipt of the expected return on that investment is risk neutral. A risk-loving consumer would prefer a risky investment with a given expected return to the certain receipt of that expected return.
5. Risk can be reduced by (a) diversification, (b) insurance, and (c) additional information.
6. The *law of large numbers* enables insurance companies to provide insurance for which the premiums paid equal the expected value of the losses being insured against. We call such insurance *actuarially fair*.
7. Consumer theory can be applied to decisions to invest in risky assets. The budget line reflects the price of risk, and consumers' indifference curves reflect their attitudes toward risk.
8. Individual behavior sometimes seems unpredictable, even irrational, and contrary to the assumptions that underlie the basic model of consumer choice. The study of behavioral economics enriches consumer theory by accounting for *reference points*, *endowment effects*, *anchoring*, fairness considerations, and deviations from the laws of probability.

QUESTIONS FOR REVIEW

1. What does it mean to say that a person is *risk averse*? Why are some people likely to be risk averse while others are risk lovers?
2. Why is the variance a better measure of variability than the range?
3. George has \$5000 to invest in a mutual fund. The expected return on mutual fund A is 15 percent and the expected return on mutual fund B is 10 percent. Should George pick mutual fund A or fund B?
4. What does it mean for consumers to maximize expected utility? Can you think of a case in which a person might *not* maximize expected utility?
5. Why do people often want to insure fully against uncertain situations even when the premium paid exceeds the expected value of the loss being insured against?
6. Why is an insurance company likely to behave as if it were risk neutral even if its managers are risk-averse individuals?

³⁵Henry S. Farber, "Is Tomorrow Another Day? The Labor Supply of New York City Cabdrivers," *Journal of Political Economy* 113 (2005): 46–82.

³⁶See Vincent P. Crawford and Juanjuan Meng, "New York City Cab Drivers' Labor Supply Revisited: Reference-Dependent Preferences with Rational-Expectations Targets for Hours and Income," *American Economic Review*, 101 (August 2011): 1912–1934.



7. When is it worth paying to obtain more information to reduce uncertainty?
8. How does the diversification of an investor's portfolio avoid risk?
9. Why do some investors put a large portion of their portfolios into risky assets while others invest largely in risk-free alternatives? (*Hint: Do the two investors receive exactly the same return on average? If so, why?*)
10. What is an endowment effect? Give an example of such an effect.
11. Jennifer is shopping and sees an attractive shirt. However, the price of \$50 is more than she is willing to pay. A few weeks later, she finds the same shirt on sale for \$25 and buys it. When a friend offers her \$50 for the shirt, she refuses to sell it. Explain Jennifer's behavior.

EXERCISES

1. Consider a lottery with three possible outcomes:
 - \$125 will be received with probability .2
 - \$100 will be received with probability .3
 - \$50 will be received with probability .5
 - a. What is the expected value of the lottery?
 - b. What is the variance of the outcomes?
 - c. What would a risk-neutral person pay to play the lottery?
2. Suppose you have invested in a new computer company whose profitability depends on two factors: (1) whether the U.S. Congress passes a tariff raising the cost of Japanese computers and (2) whether the U.S. economy grows slowly or quickly. What are the four mutually exclusive states of the world that you should be concerned about?
3. Richard is deciding whether to buy a state lottery ticket. Each ticket costs \$1, and the probability of winning payoffs is given as follows:

PROBABILITY	RETURN
.5	\$0.00
.25	\$1.00
.2	\$2.00
.05	\$7.50

- a. What is the expected value of Richard's payoff if he buys a lottery ticket? What is the variance?
 - b. Richard's nickname is "No-Risk Rick" because he is an extremely risk-averse individual. Would he buy the ticket?
 - c. Richard has been given 1000 lottery tickets. Discuss how you would determine the smallest amount for which he would be willing to sell all 1000 tickets.
 - d. In the long run, given the price of the lottery tickets and the probability/return table, what do you think the state would do about the lottery?
4. Suppose an investor is concerned about a business choice in which there are three prospects—the probability and returns are given below:

PROBABILITY	RETURN
.4	\$100
.3	30
.3	−30

What is the expected value of the uncertain investment? What is the variance?

5. You are an insurance agent who must write a policy for a new client named Sam. His company, Society for Creative Alternatives to Mayonnaise (SCAM), is working on a low-fat, low-cholesterol mayonnaise substitute for the sandwich-condiment industry. The sandwich industry will pay top dollar to the first inventor to patent such a mayonnaise substitute. Sam's SCAM seems like a very risky proposition to you. You have calculated his possible returns table as follows:

PROBABILITY	RETURN	OUTCOME
.999	−\$1,000,000	(he fails)
.001	\$1,000,000,000	(he succeeds and sells his formula)

- a. What is the expected return of Sam's project? What is the variance?
 - b. What is the most that Sam is willing to pay for insurance? Assume Sam is risk neutral.
 - c. Suppose you found out that the Japanese are on the verge of introducing their own mayonnaise substitute next month. Sam does not know this and has just turned down your final offer of \$1000 for the insurance. Assume that Sam tells you SCAM is only six months away from perfecting its mayonnaise substitute *and* that you know what you know about the Japanese. Would you raise or lower your policy premium on any subsequent proposal to Sam? Based on his information, would Sam accept?
6. Suppose that Natasha's utility function is given by $u(I) = \sqrt{10I}$, where I represents annual income in thousands of dollars.



- a. Is Natasha risk loving, risk neutral, or risk averse? Explain.
 - b. Suppose that Natasha is currently earning an income of \$40,000 ($I = 40$) and can earn that income next year with certainty. She is offered a chance to take a new job that offers a .6 probability of earning \$44,000 and a .4 probability of earning \$33,000. Should she take the new job?
 - c. In (b), would Natasha be willing to buy insurance to protect against the variable income associated with the new job? If so, how much would she be willing to pay for that insurance? (*Hint: What is the risk premium?*)
7. Suppose that two investments have the same three payoffs, but the probability associated with each payoff differs, as illustrated in the table below:

PAYOFF	PROBABILITY (INVESTMENT A)	PROBABILITY (INVESTMENT B)
\$300	0.10	0.30
\$250	0.80	0.40
\$200	0.10	0.30

- a. Find the expected return and standard deviation of each investment.
 - b. Jill has the utility function $U = 5I$, where I denotes the payoff. Which investment will she choose?
 - c. Ken has the utility function $U = 5\sqrt{I}$. Which investment will he choose?
 - d. Laura has the utility function $U = 5I^2$. Which investment will she choose?
8. As the owner of a family farm whose wealth is \$250,000, you must choose between sitting this season out and investing last year's earnings (\$200,000) in a safe money market fund paying 5.0 percent or planting summer corn. Planting costs \$200,000, with a six-month time to harvest. If there is rain, planting summer corn will yield \$500,000 in revenues at harvest. If there is a drought, planting will yield \$50,000 in revenues. As a third choice, you can purchase AgriCorp drought-resistant summer corn at a cost of \$250,000 that will yield \$500,000 in revenues at harvest if there is rain, and \$350,000 in revenues if there is a drought. You are risk averse, and your preference for family wealth (W) is specified by the relationship $U(W) = \sqrt{W}$. The probability of a summer drought is 0.30, while the probability of summer rain is 0.70.
- Which of the three options should you choose? Explain.
9. Draw a utility function over income $u(I)$ that describes a man who is a risk lover when his income is low but risk averse when his income is high. Can you explain why such a utility function might reasonably describe a person's preferences?
 10. A city is considering how much to spend to hire people to monitor its parking meters. The following information is available to the city manager:
 - Hiring each meter monitor costs \$10,000 per year.
 - With one monitoring person hired, the probability of a driver getting a ticket each time he or she parks illegally is equal to .25.
 - With two monitors, the probability of getting a ticket is .5; with three monitors, the probability is .75; and with four, it's equal to 1.
 - With two monitors hired, the current fine for over-time parking is \$20.
 - a. Assume first that all drivers are risk neutral. What parking fine would you levy, and how many meter monitors would you hire (1, 2, 3, or 4) to achieve the current level of deterrence against illegal parking at the minimum cost?
 - b. Now assume that drivers are highly risk averse. How would your answer to (a) change?
 - c. (For discussion) What if drivers could insure themselves against the risk of parking fines? Would it make good public policy to permit such insurance?
 11. A moderately risk-averse investor has 50 percent of her portfolio invested in stocks and 50 percent in risk-free Treasury bills. Show how each of the following events will affect the investor's budget line and the proportion of stocks in her portfolio:
 - a. The standard deviation of the return on the stock market increases, but the expected return on the stock market remains the same.
 - b. The expected return on the stock market increases, but the standard deviation of the stock market remains the same.
 - c. The return on risk-free Treasury bills increases.
 12. Suppose there are two types of e-book consumers: 100 "standard" consumers with demand $Q = 20 - P$ and 100 "rule of thumb" consumers who buy 10 e-books only if the price is less than \$10. (Their demand curve is given by $Q = 10$ if $P < 10$ and $Q = 0$ if $P \geq 10$.) Draw the resulting total demand curve for e-books. How has the "rule of thumb" behavior affected the elasticity of total demand for e-books?

CHAPTER 8

An Overview of the Financial System

LEARNING OBJECTIVES

After studying this chapter you should be able to

1. summarize the basic function performed by financial markets
2. explain why financial markets are classified as debt and equity markets, primary and secondary markets, exchanges and over-the-counter markets, and money and capital markets
3. describe the principal money market and capital market instruments
4. express why the government regulates financial markets and financial intermediaries (i.e., chartered banks, trust and mortgage loan companies, credit unions and *caisses pop laires*, insurance companies, mutual fund companies, and other institutions)

PREVIEW

Inez the Inventor has designed a low-cost robot that cleans house (even does windows), washes the car, and mows the lawn, but she has no funds to put her wonderful invention into production. Walter the Widower has plenty of savings, which he and his wife accumulated over the years. If Inez and Walter could get together so that Walter could provide funds to Inez, Inez's robot would see the light of day, and the economy would be better off: we would have cleaner houses, shinier cars, and more beautiful lawns.

Financial markets (bond and stock markets) and financial intermediaries (banks, insurance companies, pension funds) have the basic function of getting people like Inez and Walter together by moving funds from those who have a surplus of funds (Walter) to those who have a shortage of funds (Inez). More realistically, when Apple invents a better iPod, it may need funds to bring it to market. Similarly, when a local government needs to build a road or a school, it may need more funds than local property taxes provide. Well-functioning financial markets and financial intermediaries are crucial to economic health.

To study the effects of financial markets and financial intermediaries on the economy, we need to acquire an understanding of their general structure and operation.

In this chapter we learn about the major financial intermediaries and the instruments that are traded in financial markets as well as how these markets are regulated.

This chapter presents an overview of the fascinating study of financial markets and institutions. We return to a more detailed treatment of the regulation, structure, and evolution of financial markets in Chapters 8 through 12.

FUNCTION OF FINANCIAL MARKETS

Financial markets perform the essential economic function of channelling funds from households, firms, and governments who have saved surplus funds by spending less than their income to those who have a shortage of funds because they wish to spend more than they earn. This function is shown schematically in Figure 2-1. Those who have saved and are lending funds, the lender-savers, are at the left, and those who must borrow funds to finance their spending, the borrower-spenders, are at the right. The principal lender-savers are households, but business enterprises and the government (particularly provincial and local government), as well as foreigners and their governments, sometimes also find themselves with excess funds and so lend them out. The most important borrower-spenders are businesses and the government (particularly the federal government), but households and foreigners also borrow to finance their purchases of cars, furniture, and houses. The arrows show that funds flow from lender-savers to borrower-spenders via two routes.

In *direct finance* (the route at the bottom of Figure 2-1), borrowers borrow funds directly from lenders in financial markets by selling them *securities* (also called *financial instruments*), which are claims on the borrower's future income or assets. Securities are assets for the person who buys them but **liabilities** (IOUs or debts) for the individual or firm that sells (issues) them. For example, if Research In Motion (RIM) needs to borrow funds to pay for a new factory to manufacture new products, it might borrow the funds from savers by selling them *bonds*, debt securities that promise to make payments periodically for a specified period of time, or *stocks*, securities that entitle the owners to a share of the company's profits and assets.

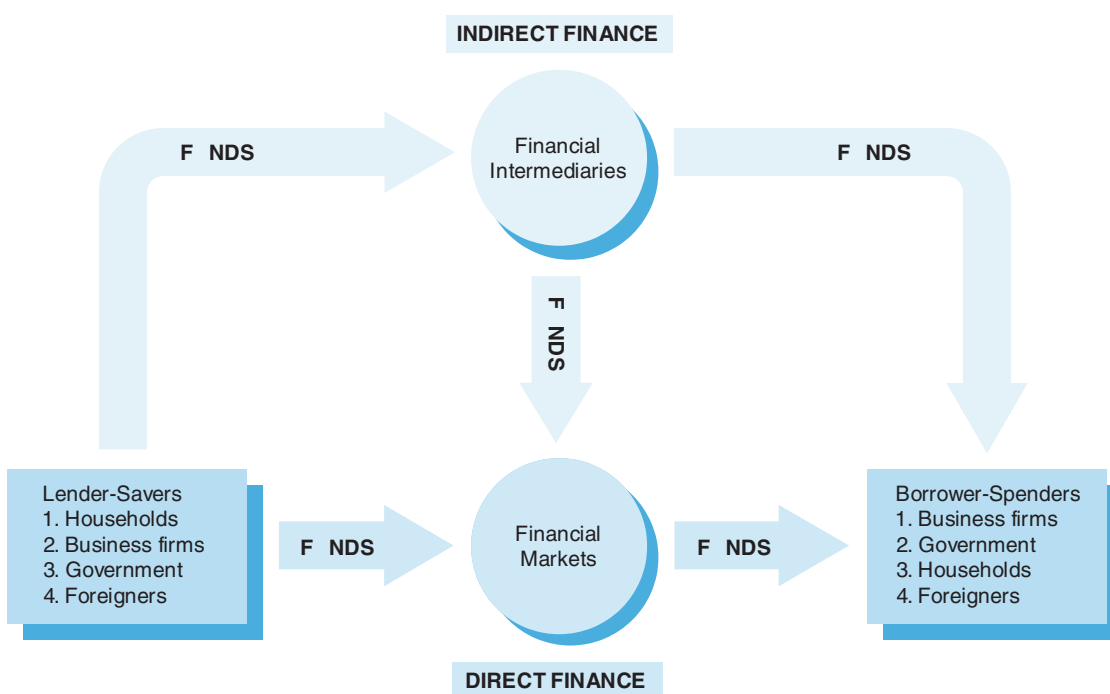


FIG RE 8-1 Flows of Funds Through the Financial System

Why is this channelling of funds from savers to spenders so important to the economy? The answer is that the people who save are frequently not the same people who have profitable investment opportunities available to them, the entrepreneurs. Let's first think about this on a personal level. Suppose that you have saved \$1000 this year, but no borrowing or lending is possible because there are no financial markets. If you do not have an investment opportunity that will permit you to earn income with your savings, you will just hold on to the \$1000 and will earn no interest. However, Carl the Carpenter has a productive use for your \$1000: he can use it to purchase a new tool that will shorten the time it takes him to build a house, thereby earning him an extra \$200 per year. If you could get in touch with Carl, you could lend him the \$1000 at a rental fee (interest) of \$100 per year, and both of you would be better off. You would earn \$100 per year on your \$1000, instead of the zero amount that you would earn otherwise, while Carl would earn \$100 more income per year (the \$200 extra earnings per year minus the \$100 rental fee for the use of the funds).

In the absence of financial markets, you and Carl the Carpenter might never get together. You would both be stuck with the status quo, and both of you would be worse off. Without financial markets, it is hard to transfer funds from a person who has no investment opportunities to one who has them; financial markets are thus essential to promoting economic efficiency.

The existence of financial markets is beneficial even if someone borrows for a purpose other than increasing production in a business. Say that you are recently married, have a good job, and want to buy a house. You earn a good salary, but because you have just started to work, you have not saved much. Over time you would have no problem saving enough to buy the house of your dreams, but by then you would be too old to get full enjoyment from it. Without financial markets, you are stuck; you cannot buy the house and must continue to live in your tiny apartment.

If a financial market were set up so that people who had built up savings could lend you the funds to buy the house, you would be more than happy to pay them some interest in order to own a home while you are still young enough to enjoy it. Then, over time, you would pay back your loan. If this loan could occur, you would be better off, as would the persons who made you the loan. They would now earn some interest, whereas they would not if the financial market did not exist.

Now we can see why financial markets have such an important function in the economy. They allow funds to move from people who lack productive investment opportunities to people who have such opportunities. Financial markets are critical for producing an efficient allocation of capital, which contributes to higher production and efficiency for the overall economy. Indeed, as we will explore in Chapter 9, when financial markets break down during financial crises (as they have in Mexico, East Asia, and Argentina in recent years), severe economic hardship results, which can even lead to dangerous political instability.

Well-functioning financial markets also directly improve the well-being of consumers by allowing them to time their purchases better. They provide funds to young people to buy what they need and can eventually afford without forcing them to wait until they have saved up the entire purchase price. Financial markets that are operating efficiently improve the economic welfare of everyone in the society.

STRUCTURE OF FINANCIAL MARKETS

Now that we understand the basic function of financial markets, let's look at their structure. The following descriptions of several categorizations of financial markets illustrate essential features of these markets.

Debt and Equity Markets

A firm or an individual can obtain funds in a financial market in two ways. The most common method is to issue a debt instrument, such as a bond or a mortgage, which is a contractual agreement by the borrower to pay the holder of the instrument fixed dollar amounts at regular intervals (interest and principal payments) until a specified date (the maturity date), when a final payment is made. The **maturity** of a debt instrument is the number of years (term) until that instrument's expiration date. A debt instrument is **short-term** if its maturity is less than a year and **long-term** if its maturity is ten years or longer. Debt instruments with a maturity between one and ten years are said to be **intermediate-term**.

The second method of raising funds is by issuing **equities**, such as common stock, which are claims to share in the net income (income after expenses and taxes) and the assets of a business. If you own one share of common stock in a company that has issued one million shares, you are entitled to one-millionth of the firm's net income and one-millionth of the firm's assets. Equities often make periodic payments (**dividends**) to their holders and are considered long-term securities because they have no maturity date. In addition, owning stock means that you own a portion of the firm and thus have the right to vote on issues important to the firm and to elect its directors.

The main disadvantage of owning a corporation's equities rather than its debt is that an equity holder is a *residual claimant*; that is, the corporation must pay all its debt holders before it pays its equity holders. The advantage of holding equities is that equity holders benefit directly from any increases in the corporation's profitability or asset value because equities confer ownership rights on the equity holders. Debt holders do not share in this benefit because their dollar payments are fixed. We examine the pros and cons of debt versus equity instruments in more detail in Chapter 8, which provides an economic analysis of financial structure.

Primary and Secondary Markets

A **primary market** is a financial market in which new issues of a security, such as a bond or a stock, are sold to initial buyers by the corporation or government agency borrowing the funds. A **secondary market** is a financial market in which securities that have been previously issued can be resold.

The primary markets for securities are not well known to the public because the selling of securities to initial buyers often takes place behind closed doors. An important financial institution that assists in the initial sale of securities in the primary market is the **investment bank**. It does this by **underwriting** securities: it guarantees a price for a corporation's securities and then sells them to the public.

The Toronto Stock Exchange (TSX) and the TSX Venture Exchange, in which previously issued stocks are traded, are the best-known examples of Canadian secondary markets, although the bond markets, in which previously issued bonds of major corporations and the Canadian government are bought and sold, actually have a larger trading volume. Other examples of secondary markets are foreign exchange markets, futures markets, and options markets. Securities brokers and dealers are crucial to a well-functioning secondary market. **Brokers** are agents of investors who match buyers with sellers of securities; **dealers** link buyers and sellers by buying and selling securities at stated prices.

When an individual buys a security in the secondary market, the person who has sold the security receives money in exchange for the security, but the corporation that issued the security acquires no new funds. A corporation acquires new funds only when its securities are first sold in the primary market. Nonetheless, secondary markets serve two important functions. First, they make it easier to sell these financial instruments to raise cash; that is, they make the financial instruments more **liquid**. The increased liquidity of these instruments then makes them more desirable and thus easier for the issuing firm to sell in the primary market. Second, they determine the price of the security that the issuing firm sells in the primary market. The investors that buy securities in the primary market will pay the issuing corporation no more than the price they think the secondary market will set for this security. The higher the security's price in the secondary market, the higher will be the price that the issuing firm will receive for a new security in the primary market and hence the greater the amount of financial capital it can raise. Conditions in the secondary market are therefore the most relevant to corporations issuing securities. It is for this reason that books like this one, which deal with financial markets, focus on the behaviour of secondary markets rather than that of primary markets.

Exchanges and Over-the-Counter Markets

Secondary markets can be organized in two ways. One is to organize **exchanges**, where buyers and sellers of securities (or their agents or brokers) meet in one central location to conduct trades. The Toronto Stock Exchange for stocks and the Winnipeg Commodity Exchange for commodities (wheat, oats, barley, and other agricultural commodities) are examples of organized exchanges. The Montreal Exchange (ME) is another example of an organized exchange, offering a range of equity, interest rate, and index derivative products.

The other method of organizing a secondary market is to have an **over-the-counter (OTC) market**, in which dealers at different locations who have an inventory of securities stand ready to buy and sell securities “over the counter” to anyone who comes to them and is willing to accept their prices. Because over-the-counter dealers are in computer contact and know the prices set by one another, the OTC market is very competitive and not very different from a market with an organized exchange.

Many common stocks are traded over the counter, although a majority of the largest corporations have their shares traded at organized stock exchanges. The Canadian government bond market, by contrast, is set up as an over-the-counter market. Dealers establish a “market” in these securities by standing ready to buy and sell Canadian government bonds. Other over-the-counter markets include those that trade other types of financial instruments such as negotiable certificates of deposit, overnight funds, and foreign exchange.

Money and Capital Markets

Another way of distinguishing between markets is on the basis of the maturity of the securities traded in each market. The **money market** is a financial market in which only short-term debt instruments (generally those with original maturity of less than one year) are traded; the **capital market** is the market in which longer-term debt (generally those with original maturity of one year or greater) and equity instruments are traded. Money market securities are usually more widely traded than longer-term securities and so tend to be more liquid. In addition, as we will see in Chapter 4, short-term securities have smaller fluctuations in prices than long-term securities, making them safer investments. As a result, corporations and

banks actively use the money market to earn interest on surplus funds that they expect to have only temporarily. Capital market securities, such as stocks and long-term bonds, are often held by financial intermediaries such as insurance companies and pension funds, which have more certainty about the amount of funds they will have available in the future.

FINANCIAL MARKET INSTRUMENTS

To complete our understanding of how financial markets perform the important role of channelling funds from lender-savers to borrower-spenders, we need to examine the securities (instruments) traded in financial markets. We first focus on the instruments traded in the money market and then turn to those traded in the capital market.

Money Market Instruments

Because of their short terms to maturity, the debt instruments traded in the money market undergo the least price fluctuations and so are the least risky investments. The money market has undergone great changes in the past three decades, with the amount of some financial instruments growing at a far more rapid rate than others.

The principal money market instruments are listed in Table 2-1, along with the amount outstanding at the end of 1980, 1990, 2000, and 2008. The *National Post: Financial Post* reports money market rates in its “Bond Yields and Rates” column (see the Financial News: Money Rates box on page 24).

GOVERNMENT OF CANADA TREASURY BILLS These short-term debt instruments of the Canadian government are issued in 1-, 3-, 6-, and 12-month maturities to finance the federal government. They pay a set amount at maturity and have no interest payments, but they effectively pay interest by initially selling at a discount, that is, at a price lower than the set amount paid at maturity. For instance, you might pay \$9600 in May 2010 for a one-year treasury bill that can be redeemed in May 2011 for \$10 000.

Treasury bills are the most liquid of all the money market instruments because they are the most actively traded. They are also the safest of all money market instruments because there is almost no possibility of *default*, a situation in which the party issuing the debt instrument (the federal government, in this case) is unable to make interest payments or pay off the amount owed when the instrument matures. The federal government is always able to meet its debt obligations, because it can raise taxes to pay off its debts. Treasury bills are held mainly by banks, although households, corporations, and other financial intermediaries hold small amounts.

TABLE 8-1 Principal Money Market Instruments

Type of Instrument	Amount Outstanding (\$ millions)			
	1980	1990	2000	2008
Treasury bills				
Government of Canada	13 709	113 654	76 633	116 706
Provincial governments	905	12 602	17 541	24 646
Municipal governments	113	514	188	155
Short-term paper				
Commercial paper	2 555	12 971	24 330	13 063

Source: Statistics Canada CANSIM II series V37377, V122256, V122257, and V122652.

CERTIFICATES OF DEPOSIT A *certificate of deposit* (CD) is a debt instrument sold by a bank to depositors that pays annual interest of a given amount and at maturity pays back the original purchase price. CDs are often negotiable, meaning that they can be traded, and in bearer form (called **bearer deposits**), meaning that the buyer's name is neither recorded in the issuer's books nor on the security itself. These negotiable CDs are issued in multiples of \$100 000 and with maturities of 30 to 365 days, and can be resold in a secondary market, thus offering the purchaser both yield and liquidity.

Chartered banks also issue non-negotiable CDs. That is, they cannot be sold to someone else and cannot be redeemed from the bank before maturity without paying a substantial penalty. Non-negotiable CDs are issued in denominations ranging from \$5000 to \$100 000 and with maturities of one day to five years. They are also known as **term deposits** or **term notes**.

CDs are also an extremely important source of funds for trust and mortgage loan companies. These institutions issue CDs under a variety of names; for example, DRs (Deposit Receipts), GTCs (Guaranteed Trust Certificates), GICs (Guaranteed Investment Certificates), and GIRs (Guaranteed Investment Receipts).

COMMERCIAL PAPER *Commercial paper* is an unsecured short-term debt instrument issued in either Canadian dollars or other currencies by large banks and well-known corporations, such as Microsoft and Bombardier. Because commercial paper is unsecured, only the largest and most creditworthy corporations issue commercial paper. The interest rate the corporation is charged reflects the firm's level of risk. The interest rate on commercial paper is low relative to those on other corporate fixed-income securities and slightly higher than rates on government of Canada treasury bills.

Sales finance companies also issue short-term promissory notes known as **finance paper**. Finance and commercial paper are issued in minimum denominations of \$50 000 and in maturities of 30 to 365 days for finance paper and 1 to 365 days for commercial paper. Most finance and commercial paper is issued on a discounted basis. Chapter 11 discusses why the commercial paper market has had such tremendous growth.

REPURCHASE AGREEMENTS *Repurchase agreements*, or *repos*, are effectively short-term loans (usually with a maturity of less than two weeks) for which treasury bills serve as *collateral*, an asset that the lender receives if the borrower does not pay back the loan. Repos are made as follows: a large corporation, such as Bombardier, may have some idle funds in its bank account, say \$1 million, which it would like to lend for a week. Bombardier uses this excess \$1 million to buy treasury bills from a bank, which agrees to repurchase them the next week at a price slightly above Bombardier's purchase price. The effect of this agreement is that Bombardier makes a loan of \$1 million to the bank and holds \$1 million of the bank's treasury bills until the bank repurchases the bills to pay off the loan. Repurchase agreements are a fairly recent innovation in financial markets, having been introduced in 1969. They are now an important source of bank funds, with the most important lenders in this market being large corporations.

OVERNIGHT FUNDS These are typically overnight loans by banks to other banks. The *overnight funds* designation is somewhat confusing, because these loans are not made by the federal government or by the Bank of Canada, but rather by banks to other banks. One reason why a bank might borrow in the overnight funds market is that it might find it does not have enough settlement

FINANCIAL NEWS

Money Rates

The *Globe and Mail* and the *National Post* publish daily a listing of interest rates on many different financial instruments. In the *National Post: Financial Post*, this listing can be found in the “Bond Yields and Rates” column.

The interest rates in the “Bond Yields and Rates” column that are discussed most frequently in the media are as follows:

Bank rate: The interest rate charged by the Bank of Canada on loans made to members of the Canadian Payments Association.

Overnight money market (financing) rate: A measure of the collateralized overnight rate compiled by the Bank of Canada.

Prime rate: The base interest rate on corpo-

rate bank loans, an indicator of the cost of business borrowing from banks.

Treasury bill rates: The interest rates on Government of Canada treasury bills, an indicator of general interest rate movements.

Selected U.S. interest rates: Selected U.S. interest rates such as treasury bill rates, commercial paper rates, the discount rate, the prime rate, and the federal funds rate. These are indicators of general interest rate movements in the United States.

London interbank offer rate (or Libor): The British Bankers’ Association average of interbank rates for dollar deposits in the London market.

BOND YIELDS AND RATES

CANADIAN YIELDS

	Latest	Prev day	Wk ago	4 Wks ago
T-Bills				
1-month	0.30	0.27	0.49	0.53
3-month	0.40	0.43	0.60	0.72
6-month	0.49	0.49	0.65	0.72
1-year	0.55	0.57	0.75	0.80

Bonds

2-year	0.95	0.93	1.14	1.14
5-year	1.85	1.81	2.05	1.93
7-year	2.07	2.03	2.25	2.36
10-year	2.93	2.91	3.12	3.04
30-year	3.61	3.55	3.70	3.76

Banker's acceptances (ask prices)

1-month	0.62	0.65	0.75	0.95
3-month	0.61	0.62	0.75	0.90
6-month	0.80	0.85	1.10	1.30

3-mth forward rate agreement

3-month	0.51	0.49	0.65	0.63
6-month	0.45	0.46	0.70	0.59
9-month	0.56	0.54	0.86	0.71

U.S. YIELDS

	Latest	Prev day	Wk ago	4 Wks ago
T-Bills				
1-month	0.05	0.08	0.13	0.21
3-month	0.19	0.19	0.24	0.28
6-month	0.38	0.37	0.43	0.42

Bonds

2-year	0.94	0.88	0.97	0.99
5-year	1.88	1.79	1.99	1.96
10-year	2.53	2.47	2.68	2.98
30-year	2.87	2.81	3.02	3.69

Commercial paper

1-month	0.24	0.21	0.27	0.34
3-month	0.35	0.34	0.51	0.43
6-month	1.48	1.58	1.58	1.28

3-mth forward rate agreement

3-month	1.43	1.32	1.28	1.07
6-month	1.44	1.36	1.34	1.15
9-month	1.57	1.49	1.50	1.35

INTERNATIONAL

	Latest	Prev day	Wk ago	4 Wks ago
Euro-deposit rates (bid)				
US\$ 1-month	0.47	0.45	0.43	1.20
3-month	1.45	1.24	1.21	1.13
6-month	1.85	1.69	1.64	1.56
C\$ 3-month	1.21	1.21	1.35	1.52
euro 3-month	1.70	1.66	1.49	1.99
Yen 3-month	0.90	0.32	0.78	0.65
£ 3-month	1.71	1.61	1.78	1.69

London interbank offer rate US\$

US\$ 1-month	n.a	0.53	0.50	0.45
3-month	n.a	1.28	1.26	1.24
6-month	n.a	1.83	1.80	1.75

BANK RATES

Canada		United States	
Bank of Canada	0.75	Discount	0.50
Overnight Money		Prime	3.25
Market Financing	0.30	Federal Funds	0.22
Prime	2.50		
Call Loan Average	0.45		

Source: *National Post: Financial Post*, March 7, 2009, p. FP13. All rights reserved. Republication or redistribution of Thomson Reuters content, including by framing or similar means, is expressly prohibited without the prior written consent of Thomson Reuters. Thomson Reuters and its logo are registered trademarks or trademarks of the Thomson Reuters group of companies around the world. © Thomson Reuters 2009. Thomson Reuters journalists are subject to an *Editorial Handbook*, which requires fair presentation and disclosure of relevant interests.

deposits at the Bank of Canada. It can then borrow these balances from another bank with excess settlement balances.

The overnight market is very sensitive to the credit needs of the deposit-taking institutions, so the interest rate on overnight loans, called the **overnight interest rate**, is a closely watched barometer of the tightness of credit market conditions

in the banking system and the stance of monetary policy. When it is high, it indicates that the banks are strapped for funds, whereas when it is low, banks' credit needs are low.

Capital Market Instruments

Capital market instruments are debt and equity instruments with maturities of greater than one year. They have far wider price fluctuations than money market instruments and are considered to be fairly risky investments. The principal capital market instruments are listed in Table 2-2, which shows the amount outstanding at the end of 1980, 1990, 2000, and 2008.

STOCKS *Stocks* are equity claims on the net income and assets of a corporation. Their value was \$324.1 billion at the end of 2008. The amount of new stock issues in any given year is typically quite small—less than 1% of the total value of shares outstanding. Individuals hold around half of the value of stocks; pension funds, mutual funds, and insurance companies hold the rest.

MORTGAGES *Mortgages* are loans to households or firms to purchase housing, land, or other real structures, where the structure or land serves as collateral for the loans. The mortgage market is the largest debt market in Canada, with the amount of residential mortgages (used to purchase residential housing) outstanding more than tenfold the amount of commercial and farm mortgages. Trust and mortgage loan companies and credit unions and *caisses populaires* were the primary lenders in the residential mortgage market until 1967. The revision of the Bank Act in 1967, however, extended the authority of chartered banks to make conventional residential mortgage loans and chartered banks entered this market very aggressively.

Banks and life insurance companies make the majority of commercial and farm mortgages. The federal government also plays an active role in the mortgage market via the Canada Mortgage and Housing Corporation (CMHC), which provides funds to the mortgage market by selling bonds and using the proceeds to buy mortgages.

CORPORATE BONDS These are long-term bonds issued by corporations with very strong credit ratings. The typical *corporate bond* sends the holder an interest payment twice a year and pays off the face value when the bond matures. Some

TABLE 8-2 Principal Capital Market Instruments

Type of Instrument	Amount Outstanding (\$ billions)			
	1980	1990	2000	2008
Corporate stocks (market value)	42.9	109.8	242.1	324.1
Residential mortgages	91.9	245.3	431.2	863.8
Corporate bonds	30.0	72.8	187.6	274.6
Government of Canada securities (marketable)	27.8	124.5	301.9	223.1
Bank commercial loans	58.7	102.7	132.0	185.0
Consumer loans	39.2	90.9	189.8	398.6
Nonresidential and farm mortgages	15.1	56.1	49.7	77.1

Source: Statistics Canada CANSIM II series V122642, V122746, V122640, V37378, V122631, V122707, V122656, V122657, V122658, V122659, V800015, and the authors' calculations.

corporate bonds, called *convertible bonds*, have the additional feature of allowing the holder to convert them into a specified number of shares of stock at any time up to the maturity date. This feature makes convertible bonds more desirable to prospective purchasers than regular bonds, and allows the corporation to reduce its interest payments, because these bonds can increase in value if the price of the stock appreciates sufficiently. Because the outstanding amount of both convertible and nonconvertible bonds for any given corporation is small, they are not nearly as liquid as other securities such as Government of Canada bonds.

Although the size of the corporate bond market is substantially smaller than that of the stock market, the volume of new corporate bonds issued each year is substantially greater than the volume of new stock issues. Thus the behaviour of the corporate bond market is probably far more important to a firm's financing decisions than the behaviour of the stock market. The principal buyers of corporate bonds are life insurance companies; pension funds and households are other large holders.

GOVERNMENT OF CANADA BONDS Intermediate-term bonds (those with initial maturities from one to ten years) and long-term bonds (those with initial maturities greater than ten years) are issued by the federal government to finance its deficit. Because they are the most widely traded bonds in Canada, they are the most liquid security traded in the capital market. They are held by the Bank of Canada, banks, households, and foreign investors.

These debt instruments are issued in either bearer or registered form and in denominations of \$1000, \$5000, \$25 000, \$100 000, and \$1 million. In the case of **registered bonds**, the name of the owner appears on the bond certificate and is also recorded at the Bank of Canada. Some issues have the additional **call** (or **redemption**) feature of allowing them to be “called” on specified notice (usually 30 to 60 days).

CANADA SAVING BONDS These are nonmarketable bonds issued by the government of Canada and sold each year from early October through to April 1. *Canada Savings Bonds* (CSBs) are floating-rate bonds, available in denominations from \$100 to \$10 000, and offered exclusively to individuals, estates, and specified trusts. They are issued as registered bonds and can be purchased from financial institutions or through payroll savings plans.

CSBs are different from all other bonds issued by the government of Canada in that they do not rise or fall in value, like other bonds do. They have the valuable option of being redeemable at face value plus accrued interest, at any time prior to maturity, by being presented at any financial institution. In October 1998 the government of Canada introduced another type of bond that is similar to CSBs—Canada Premium Bonds (CPBs). CPBs offer a slightly higher coupon rate than the CSBs, but can be redeemed only once a year, on the anniversary of the issue date and during the month after that date.

PROVINCIAL AND MUNICIPAL GOVERNMENT BONDS Provincial and municipal governments also issue bonds to finance expenditures on schools, roads, and other large programs. The securities issued by provincial governments are referred to as **provincial bonds** or **provincials** and those issued by municipal governments as **municipal bonds** or **municipals**—the securities issued by the federal government are referred to as **Canadas**. Provincials and municipals are denominated in either domestic currency or foreign currencies, mostly U.S. dollars, Swiss francs, and Japanese yen. They are mainly held by trustee pension plans, social security funds (predominantly the Canada Pension Plan), and foreigners.

GOVERNMENT AGENCY SECURITIES These are long-term bonds issued by various government agencies such as the Ontario Municipal Improvement Corporation and the Alberta Municipal Financing Corporation to assist municipalities to finance such items as mortgages, farm loans, or power-generating equipment. The provincial governments guarantee many of these securities. They function much like Canadas, provincials, and municipals and are held by similar parties.

CONSUMER AND BANK COMMERCIAL LOANS These are loans to consumers and businesses made principally by banks, but—in the case of consumer loans—also by finance companies.

INTERNATIONALIZATION OF FINANCIAL MARKETS

The growing internationalization of financial markets has become an important trend. Before the 1980s, U.S. financial markets were much larger than financial markets outside the United States, but in recent years the dominance of U.S. markets has been disappearing (see the Global box, Are U.S. Capital Markets Losing Their Edge?).

The extraordinary growth of foreign financial markets has been the result of both large increases in the pool of savings in foreign countries such as Japan and the deregulation of foreign financial markets, which has enabled foreign markets to expand their activities. Canadian corporations and banks are now more likely to tap international capital markets to raise needed funds, and Canadian investors often seek investment opportunities abroad. Similarly, foreign corporations and banks raise funds from Canadians, and foreigners have become important investors in Canada. A look at international bond markets and world stock markets will give us a picture of how this globalization of financial markets is taking place.

International Bond Market, Eurobonds, and Eurocurrencies

The traditional instruments in the international bond market are known as **foreign bonds**. Foreign bonds are sold in a foreign country and are denominated in that country's currency. For example, if the German automaker Porsche sells a bond in Canada denominated in Canadian dollars, it is classified as a foreign bond. Foreign bonds have been an important instrument in the international capital market for centuries. In fact, a large percentage of U.S. railroads built in the nineteenth century were financed by sales of foreign bonds in Britain.

A more recent innovation in the international bond market is the **Eurobond**, a bond denominated in a currency other than that of the country in which it is sold—for example, a bond issued by a Canadian corporation that is denominated in Japanese yen and sold in Germany. Currently, over 80% of the new issues in the international bond market are Eurobonds, and the market for these securities has grown very rapidly.

A variant of the Eurobond is **Eurocurrencies**, which are foreign currencies deposited in banks outside the home country. The most important of the Eurocurrencies are **Eurodollars**, which are U.S. dollars deposited in foreign banks outside the United States or in foreign branches of U.S. banks. Because these short-term deposits earn interest, they are similar to short-term Eurobonds. Canadian banks borrow Eurodollar deposits from other banks or from their own foreign branches, and Eurodollars are now an important source of funds for Canadian banks.

GLOBAL

Are U.S. Capital Markets Losing Their Edge?

Over the past few decades the United States lost its international dominance in a number of manufacturing industries, including automobiles and consumer electronics, as other countries became more competitive in global markets. Recent evidence suggests that financial markets are now undergoing a similar trend: Just as Ford and General Motors have lost global market share to Toyota and Honda, U.S. stock and bond markets recently have seen their share of sales of newly issued corporate securities slip. In 2008 the London and Hong Kong stock exchanges each handled a larger share of initial public offerings (IPO) of stock than did the New York Stock Exchange, which had been by far the dominant exchange in terms of IPO value just five years before. Likewise, the portion of new corporate bonds issued worldwide that are initially sold in U.S. capital markets has fallen below the share sold in European debt markets in each of the past two years.*

Why do corporations that issue new securities to raise capital now conduct more of this business in financial markets in Europe and Asia? Among the factors contributing to this trend are quicker adoption of technological innovation by foreign financial markets, tighter immigration controls in the United States following the terrorist attacks in 2001, and perceptions that listing on American

exchanges will expose foreign securities issuers to greater risks of lawsuits. Many people see burdensome financial regulation as the main cause, however, and point specifically to the Sarbanes-Oxley Act of 2002. The U.S. Congress passed this act after a number of accounting scandals involving U.S. corporations and the accounting firms that audited them came to light. Sarbanes-Oxley aims to strengthen the integrity of the auditing process and the quality of information provided in corporate financial statements. The costs to corporations of complying with the rules and procedures are high, especially for smaller firms, but largely avoidable if firms choose to issue their securities in financial markets outside the United States. For this reason, there is much support for revising Sarbanes-Oxley to lessen its alleged harmful effects and induce more securities issuers back to United States financial markets. However, there is not conclusive evidence to support the view that Sarbanes-Oxley is the main cause of the relative decline of U.S. financial markets and therefore in need of reform.

Discussion of the relative decline of U.S. financial markets and debate about the factors that are contributing to it likely will continue.

* "Down on the Street," *The Economist*, November 25, 2006, pp. 69–71.

Note that the name of the European currency, the euro, can create some confusion about the terms Eurobond, Eurocurrencies, and Eurodollars. A Eurobond is typically not a bond that is denominated in euros. A bond denominated in euros is called a Eurobond only if *it is sold outside the countries that have adopted the euro*. Similarly, Eurodollars have nothing to do with euros, but are instead U.S. dollars deposited in banks outside the United States.

World Stock Markets

Until recently, the U.S. stock market was by far the largest in the world, but stock markets in other countries have been growing in importance (see Table 2-3). Now the United States is not always number one: in the mid-1980s, the value of stocks

TABLE 8-3 Top 10 Stock Exchanges in the World (by Domestic Market Capitalization at Year-End 2008)

Exchange	Value (in billions of US\$)	Rank in 2008
NYSE	9 209	1
Tokyo	3 116	2
Nasdaq	2 396	3
Euronext	2 102	4
London	1 868	5
Shanghai	1 425	6
Hong Kong	1 329	7
Deutsche Börse	1 111	8
Toronto	1 033	9
BME Spanish Exchanges	948	10

Source: World Federation of Exchanges, *2008 Market Highlights*, www.world-exchanges.org/statistics.

traded in Japan had at times exceeded the value of stocks traded in the United States. The increased interest in foreign stocks has prompted the development in Canada of mutual funds that specialize in trading in foreign stock markets. Canadian investors now pay attention not only to the Canadian stock markets (the Toronto Stock Exchange and the TSX Venture Exchange) but also to stock price indexes for foreign stock markets such as the Dow Jones Industrial Average (New York), the Nikkei 225 Average (Tokyo), and the Financial Times–Stock Exchange 100-Share Index (London) (see Financial News: Foreign Stock Market Indexes).

The internationalization of financial markets is having profound effects on Canada. Foreigners not only are providing funds to corporations in Canada but also are helping finance the federal government. Without these foreign funds, the Canadian economy would have grown far less rapidly in the last twenty years. The internationalization of financial markets is also leading the way to a more integrated world economy in which flows of goods and technology between countries are more commonplace. In later chapters we will encounter many examples of the important roles that international factors play in our economy.

FUNCTION OF FINANCIAL INTERMEDIARIES: INDIRECT FINANCE

As shown in Figure 2-1 (page 18), funds can move from lenders to borrowers by a second route, called *indirect finance* because it involves a financial intermediary that stands between the lender-savers and the borrower-spenders and helps transfer funds from one to the other. A financial intermediary does this by borrowing funds from the lender-savers and then using these funds to make loans to borrower-spenders. For example, a bank might acquire funds by issuing a liability to the public (an asset for the public) in the form of savings deposits. It might then use the funds to acquire an asset by making a loan to Canadian Pacific or by buying a Canadian Pacific bond in the financial market. The ultimate result is that funds have been transferred from the public (the lender-savers) to Canadian Pacific (the borrower-spender) with the help of the financial intermediary (the bank).

FINANCIAL NEWS

Foreign Stock Market Indexes

Foreign stock market indexes are published daily in the financial pages of newspapers and on the web. The entries from Bloomberg, shown here, are explained in the text.

The first column identifies the market index; for example, the shaded entry for the S&P/TSX Composite Index. The second column, "Value,"

gives the closing value of the index, which was 11 508.53 for the S&P/TSX Composite on December 7, 2009. The "Change" column indicates the change in the index, -2.27. The "% Change" column indicates the percentage change in the index, -0.02%.

WORLD INDEXES**NORTH/LATIN AMERICA**

Index	Value	Change	% Change
Dow Jones Industrial Average	10 409.38	20.48	0.20
S&P 500 Index	1 106.86	0.88	0.08
NASDAQ Composite Index	2 194.04	-0.31	-0.01
S&P/TSX Composite Index	11 508.53	-2.27	-0.02
Mexico Bolsa Index	32 111.59	6.20	0.02
Brazil Bovespa Stock Index	68 102.76	499.23	0.74

EUROPE/AFRICA/MIDDLE EAST

Index	Value	Change	% Change
DJ Euro Stoxx 50 € Pr	2 900.71	-9.62	-0.33
FTSE 100 Index	5 321.87	-0.49	-0.01
CAC 40 Index	3 844.65	-1.97	-0.05
DAX Index	5 793.66	-23.99	-0.41
IBEX 35 Index	12 027.90	-4.30	-0.04
FTSE MIB Index	22 839.62	-86.41	-0.38
AEX-Index	321.14	0.01	0.00
OMX Stockholm 30 Index	966.98	0.01	0.00
Swiss Market Index	6 476.82	-24.34	-0.37

ASIA PACIFIC

Index	Value	Change	% Change
Nikkei 225	10 167.60	145.01	1.45
Hang Seng Index	22 324.96	173.19	-0.77
S&P/ASX 200 Index	4 676.50	-25.70	-0.55

Source: Reprinted from Bloomberg.com with permission on December 7, 2009.

The process of indirect finance using financial intermediaries, called **financial intermediation**, is the primary route for moving funds from lenders to borrowers. Indeed, although the media focus much of their attention on securities markets, particularly the stock market, financial intermediaries are a far more important source of financing for corporations than securities markets are. This is true not only for Canada but for other industrialized countries as well (see the Global box, The Importance of Financial Intermediaries to Securities Markets: An International Comparison). Why are financial intermediaries and indirect finance so important in financial markets? To answer this question, we need to understand the role of transaction costs, risk sharing, and information costs in financial markets.

GL BAL

The Importance of Financial Intermediaries to Securities Markets: An International Comparison

Patterns of financing corporations differ across countries, but one key fact emerges. Studies of the major developed countries, including Canada, the United States, Great Britain, Japan, Italy, Germany, and France, show that when businesses go looking for funds to finance their activities, they usually obtain them indirectly through financial intermediaries and not directly from securities markets.* Even in Canada and the United States, which have the most developed securities markets in the world, loans from financial intermediaries are far more important for corporate finance than securities markets are. The countries that have made the least use of securities markets are Germany and Japan; in these two countries, financing from financial intermediaries has been almost ten times

greater than that from securities markets. However, after the deregulation of Japanese securities markets in recent years, the share of corporate financing by financial intermediaries has been declining relative to the use of securities markets.

Although the dominance of financial intermediaries over securities markets is clear in all countries, the relative importance of bond versus stock markets differs widely across countries. In the United States, the bond market is far more important as a source of corporate finance. On average, the amount of new financing raised using bonds is ten times the amount using stocks. By contrast, countries such as France and Italy make more use of equities markets than of the bond markets to raise capital.

*See, for example, Colin Mayer, "Financial Systems, Corporate Finance, and Economic Development," in *Asymmetric Information, Corporate Finance, and Investment*, ed. R. Glenn Hubbard (Chicago: University of Chicago Press, 1990), pp. 307–332.

Transaction Costs

Transaction costs, the time and money spent in carrying out financial transactions, are a major problem for people who have excess funds to lend. As we have seen, Carl the Carpenter needs \$1000 for his new tool, and you know that it is an excellent investment opportunity. You would like to lend him the funds, but to protect your investment, you have to hire a lawyer to write up the loan contract that specifies how much interest Carl will pay you, when he will make these interest payments, and when he will repay you the \$1000. Obtaining the contract will cost you \$500. When you include this transaction cost for making the loan, you realize that you can't earn enough from the deal (you spend \$500 to make perhaps \$100) and reluctantly tell Carl that he will have to look elsewhere.

This example illustrates that small savers like you or potential borrowers like Carl might be frozen out of financial markets and thus be unable to benefit from them. Can anyone come to the rescue? Financial intermediaries can.

Financial intermediaries can substantially reduce transaction costs because they have developed expertise in lowering costs and because their large size allows them to take advantage of **economies of scale**, the reduction in transaction costs per dollar of transactions as the size (scale) of transactions increases. For example, a bank knows how to find a good lawyer to produce an airtight loan contract, and this contract can be used over and over again in its loan transactions, thus lowering the legal cost per transaction. Instead of a loan contract (which may not be all that well written) costing \$500, a bank can hire a topflight lawyer for \$5000 to draw

up an airtight loan contract that can be used for 2000 loans at a cost of \$2.50 per loan. At a cost of \$2.50 per loan, it now becomes profitable for the financial intermediary to lend Carl the \$1000.

Because financial intermediaries are able to reduce transaction costs substantially, they make it possible for you to provide funds indirectly to people like Carl with productive investment opportunities. In addition, a financial intermediary's low transaction costs mean that it can provide its customers with **liquidit services**, services that make it easier for customers to conduct transactions. For example, banks provide depositors with chequing accounts that enable them to pay their bills easily. In addition, depositors can earn interest on chequing and savings accounts and yet still convert them into goods and services whenever necessary.

Risk Sharing

Another benefit made possible by the low transaction costs of financial institutions is that they can help reduce the exposure of investors to **risk**, that is, uncertainty about the returns investors will earn on assets. Financial intermediaries do this through **risk sharing**: they create and sell assets with risk characteristics that people are comfortable with, and the intermediaries then use the funds they acquire by selling these assets to purchase other assets that may have far more risk. Low transaction costs allow financial intermediaries to share risk at low cost, enabling them to earn a profit on the spread between the returns they earn on risky assets and the payments they make on the assets they have sold. This process of risk sharing is also sometimes referred to as **asset transformation**, because, in a sense, risky assets are turned into safer assets for investors.

Financial intermediaries also promote risk sharing by helping individuals to diversify and thereby lower the amount of risk to which they are exposed. **Diversification** entails investing in a collection (**portfolio**) of assets whose returns do not always move together, with the result that overall risk is lower than for individual assets. (Diversification is another name for the old adage that „you shouldn't put all your eggs in one basket.“) Low transaction costs allow financial intermediaries to do this by pooling a collection of assets into a new asset and then selling it to individuals.

Asymmetric Information: Adverse Selection and Moral Hazard

The presence of transaction costs in financial markets explains, in part, why financial intermediaries and indirect finance play such an important role in financial markets. An additional reason is that in financial markets, one party often does not know enough about the other party to make accurate decisions. This inequality is called **as mmetric information**. For example, a borrower who takes out a loan usually has better information about the potential returns and risk associated with the investment projects for which the funds are earmarked than the lender does. Lack of information creates problems in the financial system on two fronts: before the transaction is entered into and after.

Adverse selection is the problem created by asymmetric information *before* the transaction occurs. Adverse selection in financial markets occurs when the potential borrowers who are the most likely to produce an undesirable (*ad erse*) outcome the bad credit risks are the ones who most actively seek out a loan and are thus most likely to be selected. Because adverse selection makes it more likely that loans might be made to bad credit risks, lenders may decide not to make any loans even though there are good credit risks in the marketplace.

To understand why adverse selection occurs, suppose that you have two aunts to whom you might make a loan—Aunt Sheila and Aunt Louise. Aunt Louise is a conservative type who borrows only when she has an investment she is quite sure will pay off. Aunt Sheila, by contrast, is an inveterate gambler who has just come across a get-rich-quick scheme that will make her a millionaire if she can just borrow \$1000 to invest in it. Unfortunately, as with most get-rich-quick schemes, there is a high probability that the investment won't pay off and that Aunt Sheila will lose the \$1000.

Which of your aunts is more likely to call you to ask for a loan? Aunt Sheila, of course, because she has so much to gain if the investment pays off. You, however, would not want to make a loan to her because there is a high probability that her investment will turn sour and she will be unable to pay you back.

If you knew both your aunts very well—that is, if your information were not asymmetric—you wouldn't have a problem because you would know that Aunt Sheila is a bad risk and so you would not lend to her. Suppose, though, that you don't know your aunts well. You are more likely to lend to Aunt Sheila than to Aunt Louise because Aunt Sheila would be hounding you for the loan. Because of the possibility of adverse selection, you might decide not to lend to either of your aunts, even though there are times when Aunt Louise, who is an excellent credit risk, might need a loan for a worthwhile investment.

Moral hazard is the problem created by asymmetric information *after* the transaction occurs. Moral hazard in financial markets is the risk (*bazard*) that the borrower might engage in activities that are undesirable (*immoral*) from the lender's point of view because they make it less likely that the loan will be paid back. Because moral hazard lowers the probability that the loan will be repaid, lenders may decide that they would rather not make a loan.

As an example of moral hazard, suppose that you made a \$1000 loan to another relative, Uncle Melvin, who needs the money to purchase a computer so he can set up a business inputting students' term papers. Once you have made the loan, however, Uncle Melvin is more likely to slip off to the track and play the horses. If he bets on a 20-to-1 long shot and wins with your money, he is able to pay you back your \$1000 and live high off the hog with the remaining \$19 000. But if he loses, as is likely, you don't get paid back, and all he has lost is his reputation as a reliable, upstanding uncle. Uncle Melvin therefore has an incentive to go to the track because his gains (\$19 000) if he bets correctly may be much greater than the cost to him (his reputation) if he bets incorrectly. If you knew what Uncle Melvin was up to, you would prevent him from going to the track, and he would not be able to increase the moral hazard. However, because it is hard for you to keep informed about his whereabouts—that is, because information is asymmetric—there is a good chance that Uncle Melvin will go to the track and you will not get paid back. The risk of moral hazard might therefore discourage you from making the \$1000 loan to Uncle Melvin, even if you were sure that you would be paid back if he used it to set up his business.

The problems created by adverse selection and moral hazard are an important impediment to well-functioning financial markets. Again, financial intermediaries can alleviate these problems.

With financial intermediaries in the economy, small savers can provide their funds to the financial markets by lending these funds to a trustworthy intermediary, say, the Honest John Bank, which in turn lends the funds out either by making loans or by buying securities such as stocks or bonds. Successful financial intermediaries have higher earnings on their investments than small savers

because they are better equipped than individuals to screen out good from bad credit risks, thereby reducing losses due to adverse selection. In addition, financial intermediaries have high earnings because they develop expertise in monitoring the parties they lend to, thus reducing losses due to moral hazard. The result is that financial intermediaries can afford to pay lender-savers interest or provide substantial services and still earn a profit.

As we have seen, financial intermediaries play an important role in the economy because they provide liquidity services, promote risk sharing, and solve information problems, thereby allowing small savers and borrowers to benefit from the existence of financial markets. The success of financial intermediaries performing this role is evidenced by the fact that most Canadians invest their savings with them and obtain loans from them. Financial intermediaries play a key role in improving economic efficiency because they help financial markets channel funds from lender-savers to people with productive investment opportunities. Without a well-functioning set of financial intermediaries, it is very hard for an economy to reach its full potential. We will explore further the role of financial intermediaries in the economy in Part III.

TYPES OF FINANCIAL INTERMEDIARIES

We have seen why financial intermediaries play such an important role in the economy. Now we look at the principal financial intermediaries and how they perform the intermediation function. They fall into three categories: depository institutions (banks and near banks), contractual savings institutions, and investment intermediaries. Table 2-4 provides a guide to the discussion of the financial intermediaries that fit into these three categories by describing their primary liabilities (sources of funds) and assets (uses of funds). The relative size of these intermediaries in Canada is indicated in Table 2-5.

Depositor Institutions (Banks)

Depository institutions (which for simplicity we refer to as *banks* throughout this text) are financial intermediaries that accept deposits from individuals and institutions and make loans. The study of money and banking focuses special attention on this group of financial institutions because they are involved in the creation of deposits, an important component of the money supply. These institutions include chartered banks and the so-called near banks: trust and mortgage loan companies, and credit unions and *caisses populaires*.

CHARTERED BANKS These financial intermediaries raise funds primarily by issuing chequable deposits (deposits on which cheques can be written), savings deposits (deposits that are payable on demand but do not allow their owner to write cheques), and term deposits (deposits with fixed terms to maturity). They then use these funds to make commercial, consumer, and mortgage loans and to buy Canadian government securities and provincial and municipal bonds. There are 73 chartered banks in Canada, and as a group they are the largest financial intermediary and have the most diversified portfolios (collections) of assets.

TRUST AND MORTGAGE LOAN COMPANIES (TMLs) These depository institutions, numbering 70, obtain funds primarily through chequable and nonchequable savings deposits, term deposits, guaranteed investment certificates, and debentures. In the past, these institutions were constrained in their activities and mostly made mortgage loans for residential housing. Over time, these restrictions have been

TABLE 8-4 Primary Assets and Liabilities of Financial Intermediaries

Type of Intermediary	Primary Liabilities (Sources of Funds)	Primary Assets (Uses of Funds)
Depository Institutions (Banks)		
Chartered banks	Deposits	Loans, mortgages, government bonds
Trust and mortgage loan companies	Deposits	Mortgages
Credit unions and <i>caisses populaires</i>	Deposits	Mortgages
Contractual Savings Institutions		
Life insurance companies	Premiums from policies	Corporate bonds and mortgages
Property and casualty insurance companies	Premiums from policies	Corporate bonds and stocks
Pension funds	Retirement contributions	Corporate bonds and stocks
Investment Intermediaries		
Finance companies	Finance paper, stock, bonds	Consumer and business loans
Mutual funds	Shares	Stocks and bonds
Money market mutual funds	Shares	Money market instruments

loosened so that the distinction between these depository institutions and chartered banks has blurred. These intermediaries have become more alike and are now more competitive with each other.

CREDIT UNIONS AND CAISSES POPULAIRES (CUCPs) Credit unions and *caisses populaires* are very small cooperative lending institutions organized around a particular group: union members, employees of a particular firm, and so forth. They acquire funds from deposits and primarily make mortgage and consumer loans.

Contractual Savings Institutions

Contractual savings institutions, such as insurance companies and pension funds, are financial intermediaries that acquire funds at periodic intervals on a contractual basis. Because they can predict with reasonable accuracy how much they will have to pay out in benefits in the coming years, they do not have to worry as much as depository institutions about losing funds. As a result, the liquidity of assets is not as important a consideration for them as it is for depository institutions, and they tend to invest their funds primarily in long-term securities such as corporate bonds, stocks, and mortgages.

LIFE INSURANCE COMPANIES Life insurance companies, numbering 94, insure people against financial hardships following a death and sell annuities (annual income payments upon retirement). They acquire funds from the premiums that people pay to keep their policies in force and use them mainly to buy corporate bonds and mortgages. They also purchase stocks but are restricted in the amount

TABLE 8-5 Relative Shares of Financial Institutions and Pension Plans Regulated by OSFI (as of March 31, 2008)

Type of Intermediary	Number	Total assets (\$ millions)	Percent (%)
Chartered Banks			
Domestic	20	2 596 712	67.92
Foreign bank subsidiaries	24	139 523	3.65
Foreign bank branches	29	79 191	2.07
Trust and Loan Companies			
Bank-owned	31	243 163	6.36
Other	39	23 292	0.61
Cooperative Credit Associations	8	21 152	0.55
Life Insurance Companies			
Canadian-incorporated	46	456 440	11.94
Foreign branches	48	15 275	0.40
Fraternal Benefit Societies			
Canadian-incorporated	10	5 809	0.15
Foreign branches	8	1 775	0.05
Property and Casualty Insurance Companies			
Canadian-incorporated	96	78 256	2.05
Foreign branches	100	30 873	0.81
Pension Plans	1 350	131 765	3.44
Total		3 823 226	100.00

Source: Office of the Superintendent of Financial Institutions Canada (OSFI), 2007–2008 Annual Report.

that the can hold. Currently, with about \$472 billion of assets, the are among the largest of the contractual savings institutions.

PROPERTY AND CASUALTY (P&C) INSURANCE COMPANIES These companies, numbering 196, insure their policy holders against loss from theft, fire, and accidents. They are very much like life insurance companies, receiving funds through premiums for their policies, but they have a greater possibility of loss of funds if major disasters occur. For this reason, they use their funds to buy more liquid assets than life insurance companies do. Their largest holding of assets is government bonds and debentures; they also hold corporate bonds and stocks.

PENSION FUNDS AND GOVERNMENT RETIREMENT FUNDS Private pension funds and provincial and municipal retirement funds provide retirement income in the form of annuities to employees who are covered by a pension plan. Funds are acquired by contributions from employers and/or from employees, who either have a contribution automatically deducted from their pay cheques or contribute voluntarily. The largest asset holdings of pension funds are corporate bonds and stocks. The establishment of pension funds has been actively encouraged by the federal government both through legislation requiring pension plans and through tax incentives to encourage contributions.

Investment Intermediaries

This category of financial intermediaries includes finance companies, mutual funds, and money market mutual funds.

FINANCE COMPANIES Finance companies raise funds by selling commercial paper (a short-term debt instrument) and by issuing stocks and bonds. They lend these funds to consumers, who make purchases of such items as furniture, automobiles, and home improvements, and to small businesses. Some finance companies are organized by a parent corporation to help sell its product. For example, Ford Credit makes loans to consumers who purchase Ford automobiles.

MUTUAL FUNDS These financial intermediaries acquire funds by selling shares to many individuals and use the proceeds to purchase diversified portfolios of stocks and bonds. Mutual funds allow shareholders to pool their resources so that they can take advantage of lower transaction costs when buying large blocks of stocks or bonds. In addition, mutual funds allow shareholders to hold more diversified portfolios than they otherwise would. Shareholders can sell (redeem) shares at any time, but the value of these shares will be determined by the value of the mutual fund's holdings of securities. Because these fluctuate greatly, the value of mutual fund shares will too; therefore, investments in mutual funds can be risky.

MONEY MARKET MUTUAL FUNDS These financial institutions have the characteristics of mutual funds but also function to some extent as depository institutions because they offer deposit-type accounts. Like most mutual funds, they sell shares to acquire funds that are then used to buy money market instruments that are both safe and very liquid. The interest on these assets is paid out to shareholders.

REGULATION OF THE FINANCIAL SYSTEM

The financial system is among the most heavily regulated sectors of the Canadian economy. The government regulates financial markets for three main reasons: to increase the information available to investors, to ensure the soundness of the financial system, and to improve control of monetary policy. We will examine how these three reasons have led to the present regulatory environment. As a study aid, the principal regulatory agencies of the Canadian financial system are listed in Table 2-6.

Increasing Information Available to Investors

Asymmetric information in financial markets means that investors may be subject to adverse selection and moral hazard problems that may hinder the efficient operation of financial markets. Risky firms or outright crooks may be the most eager to sell securities to unwary investors, and the resulting adverse selection problem may keep investors out of financial markets. Furthermore, once an investor has bought a security, thereby lending money to a firm, the borrower may have incentives to engage in risky activities or to commit outright fraud. The presence of this moral hazard problem may also keep investors away from financial markets. Government regulation can reduce adverse selection and moral hazard problems in financial markets and increase their efficiency by increasing the amount of information available to investors.

Provincial securities commissions, the most significant being the Ontario Securities Commission (OSC), administer provincial acts requiring corporations issuing securities to disclose certain information about their sales, assets, and earn-

TABLE 8-6 Principal Regulatory Agencies of the Canadian Financial System

Regulator	Agenc	Subject of Regulation	Nature of Regulations
Provincial securities and exchange commissions		Organized exchanges and financial markets	Requires disclosure of information and restrict insider trading
Bank of Canada		Chartered banks, TMLs, and CUCPs	Examines the books of the deposit-taking institutions and coordinates with the federal agencies that are responsible for financial institution regulation: OSFI and CDIC
Office of the Superintendent of Financial Institutions Canada (OSFI)		All federally regulated chartered banks, TMLs, CUCPs, life insurance companies, P&C insurance companies, and pension plans	Sets capital adequacy, accounting, and board-of-directors responsibility standards; conducts bank audits and coordinates with provincial securities commissions
Canada Deposit Insurance Corporation (CDIC)		Chartered banks, TMLs, CUCPs	Provides insurance of up to \$100 000 for each depositor at a bank, examines the books of insured banks, and imposes restrictions on assets they can hold
Québec Deposit Insurance Board		TMLs and credit cooperatives in Québec	Similar role as the CDIC
Canadian Life and Health Insurance Compensation Corporation (CompCorp)		Life insurance companies	Compensates policyholders if the issuing life insurance company goes bankrupt
P&C Insurance Compensation Corporation (PACIC)		Property and casualty insurance companies	Compensates policyholders if the issuing P&C insurance company goes bankrupt

ings to the public and restrict trading by the largest stockholders in the corporation. By requiring disclosure of this information and by discouraging insider trading, which could be used to manipulate security prices, regulators hope that investors will be better informed and be protected from abuses in financial markets. Indeed, in recent years, the OSC has been particularly active in prosecuting people involved in insider trading in Canada's largest stock exchange, the Toronto Stock Exchange (TSX).

Ensuring the Soundness of Financial Intermediaries

Asymmetric information can lead to the widespread collapse of financial intermediaries, referred to as a **financial panic**. Because providers of funds to financial intermediaries may not be able to assess whether the institutions holding their funds are sound or not, if they have doubts about the overall health of financial intermediaries they may want to pull their funds out of both sound and unsound institutions. The

possible outcome is a financial panic that produces large losses for the public and causes serious damage to the economy. To protect the public and the economy from financial panics, the government has implemented various types of regulations.

RESTRICTIONS ON ENTR Provincial banking and insurance commissions, the Bank of Canada, and the Office of the Superintendent of Financial Institutions (OSFI), an agency of the federal government, have created tight regulations governing who is allowed to set up a financial intermediary. Individuals or groups that want to establish a financial intermediary, such as a bank or an insurance company, must obtain a charter from the provincial or federal government. Only if they are upstanding citizens with impeccable credentials and a large amount of initial funds will they be given a charter.

DISCLOSURE There are stringent reporting requirements for financial intermediaries. Their bookkeeping must follow certain strict principles, their books are subject to periodic inspection, and they must make certain information available to the public.

RESTRICTIONS ON ASSETS AND ACTIVITIES There are restrictions on what financial intermediaries are allowed to do and what assets they can hold. Before you put your funds into a chartered bank or some other such institution, you would want to know that your funds are safe and that the financial intermediary will be able to meet its obligations to you. One way of doing this is to restrict the financial intermediary from engaging in certain risky activities. Another way to limit a financial intermediary's risk behaviour is to restrict it from holding certain risky assets, or at least from holding a greater quantity of these risky assets than is prudent. For example, chartered banks and other depository institutions are not allowed to hold common stock because stock prices experience substantial fluctuations. Insurance companies are allowed to hold common stock, but their holdings cannot exceed a certain fraction of their total assets.

DEPOSIT INSURANCE The government can insure people's deposits so that they do not suffer any financial loss if the financial intermediary that holds these deposits fails. The most important government agency that provides this type of insurance is the Canada Deposit Insurance Corporation (CDIC), created by an act of Parliament in 1967. It insures each depositor at a member deposit-taking financial institution up to a loss of \$100 000 per account. Except for certain wholesale branches of foreign banks, credit unions, and some provincial institutions, all deposit-taking financial institutions in Canada are members of the CDIC. All CDIC members make contributions into the CDIC fund, which are used to pay off depositors in the case of a bank's failure. The Québec Deposit Insurance Board, an organization similar to CDIC and set up at the same time as CDIC, provides insurance for TMLs and credit cooperatives in Québec.

LIMITS ON COMPETITION Politicians have often declared that unbridled competition among financial intermediaries promotes failures that will harm the public. Although the evidence that competition has this effect is extremely weak, provincial and federal governments at times have imposed restrictive regulations. For example, from 1967 to 1980 the entry of foreign banks into Canadian banking was prohibited. Since 1980, however, the incorporation of foreign bank subsidiaries

has been regulated, but Canada still ranks low with respect to the degree of competition from foreign banks.

In later chapters we will look more closely at government regulation of financial markets and will see whether it has improved their functioning.

Financial Regulation Abroad

Not surprisingly, given the similarity of the economic systems here and in the United States, Japan, and the nations of Western Europe, financial regulation in these countries is similar to financial regulation in Canada. The provision of information is improved by requiring corporations issuing securities to report details about assets and liabilities, earnings, and sales of stock, and by prohibiting insider trading. The soundness of intermediaries is ensured by licensing, periodic inspection of financial intermediaries' books, and the provision of deposit insurance.

The major differences between financial regulation in Canada and abroad relate to bank regulation. In the past, for example, the United States was the only industrialized country to subject banks to restrictions on branching, which limited banks' size and restricted them to certain geographic regions. These restrictions were abolished by legislation in 1994. U.S. and Canadian banks are also the most restricted in the range of assets they may hold. Banks in other countries frequently hold shares in commercial firms; in Japan and Germany, those stakes can be sizable.

SUMMARY

1. The basic function of financial markets is to channel funds from savers who have an excess of funds to spenders who have a shortage of funds. Financial markets can do this either through direct finance, in which borrowers borrow funds directly from lenders by selling them securities, or through indirect finance, which involves a financial intermediary that stands between the lender-savers and the borrower-spenders and helps transfer funds from one to the other. This channelling of funds improves the economic welfare of everyone in the society. Because they allow funds to move from people who have no productive investment opportunities to those who have such opportunities, financial markets contribute to economic efficiency. In addition, channelling of funds directly benefits consumers by allowing them to make purchases when they need them most.
2. Financial markets can be classified as debt and equity markets, primary and secondary markets, exchanges and over-the-counter markets, and money and capital markets.
3. The principal money market instruments (debt instruments with maturities of less than one year) are treasury bills, certificates of deposit, commercial paper, repurchase agreements, overnight funds, and Eurodollars. The principal capital market instruments (debt and equity instruments with maturities greater than one year) are stocks, mortgages, corporate bonds, Canadian government securities, Canada Savings Bonds, government agency securities, provincial and municipal government bonds, and consumer and bank commercial loans.
4. An important trend in recent years is the growing internationalization of financial markets. Eurobonds, which are denominated in a currency other than that of the country in which they are sold, are now the dominant security in the international bond market. Eurodollars, which are dollars deposited in foreign banks, are an important source of funds for Canadian banks.
5. Financial intermediaries are financial institutions that acquire funds by issuing liabilities and in turn use those funds to acquire assets by purchasing securities or making loans. Financial intermediaries play an important role in the financial system because they reduce transaction costs, allow risk sharing, and solve problems created by adverse selection and moral hazard. As a result, financial intermediaries allow small savers and borrowers to benefit from the existence of financial markets, thereby increasing the efficiency of the economy.
6. The principal financial intermediaries fall into three categories: (a) banks—chartered banks, trust and mortgage loan companies, and credit unions and *caisses populaires*; (b) contractual savings institutions—life insurance companies, property and casualty

- ality insurance companies, and pension funds; and (c) investment intermediaries—finance companies, mutual funds, and money market mutual funds.
7. The government regulates financial markets and financial intermediaries for three main reasons: to increase the information available to investors, to ensure the soundness of the financial system, and to improve control of monetary policy. Regulations include requiring disclosure of information to the public, restrictions on who can set up a financial intermediary, restrictions on what assets financial intermediaries can hold, and the provision of deposit insurance.

KEY TERMS

- | | | |
|-------------------------------|-------------------------------------|---|
| adverse selection, p. 32 | exchanges, p. 21 | over-the-counter (OTC) market, p. 21 |
| asset transformation, p. 32 | finance paper, p. 23 | overnight interest rate, p. 24 |
| asymmetric information, p. 32 | financial intermediation, p. 30 | portfolio, p. 32 |
| bearer deposit notes, p. 23 | financial panic, p. 38 | primary market, p. 20 |
| brokers, p. 20 | foreign bonds, p. 27 | provincial bonds (provincials), p. 26 |
| call (redemption) p. 26 | intermediate-term, p. 20 | registered bonds, p. 26 |
| Canadas, p. 26 | investment bank, p. 20 | risk, p. 32 |
| capital market, p. 21 | liabilities, p. 18 | risk sharing, p. 32 |
| dealers, p. 20 | liquid, p. 21 | secondary market, p. 20 |
| diversification, p. 32 | liquidity services, p. 32 | short-term, p. 20 |
| dividends, p. 20 | long-term, p. 20 | term deposit receipts (term notes), p. 23 |
| economies of scale, p. 31 | maturity, p. 20 | transaction costs, p. 31 |
| equities, p. 20 | money market, p. 21 | underwriting, p. 20 |
| Eurobond, p. 27 | moral hazard, p. 33 | |
| Eurocurrencies, p. 27 | municipal bonds (municipals), p. 26 | |
| Eurodollars, p. 27 | | |

QUESTIONS

You will find the answers to the questions marked with an asterisk in the Textbook Resources section of your MyEconLab.

- *1. Why is a share of Microsoft common stock an asset for its owner and a liability for Microsoft?
2. If I can buy a car today for \$5000 and it is worth \$10 000 in extra income next year to me because it enables me to get a job as a travelling anvil seller, should I take out a loan from Larry the Loan Shark at a 90% interest rate if no one else will give me a loan? Will I be better or worse off as a result of taking out this loan? Can you make a case for legalizing loan-sharking?
- *3. Some economists suspect that one of the reasons that economies in developing countries grow so slowly is that they do not have well-developed financial markets. Does this argument make sense?
4. Describe how authority over deposit-based financial intermediaries is split among the Bank of Canada, the OSFI, and the CDIC.
- *5. “Because corporations do not actually raise any funds in secondary markets, these markets are less important to the economy than primary markets.” Comment.
6. If you suspect that a company will go bankrupt next year, which would you rather hold, bonds issued by the company or equities issued by the company? Why?
- *7. How can the adverse selection problem explain why you are more likely to make a loan to a family member than to a stranger?
8. Think of one example in which you have had to deal with the adverse selection problem.
- *9. Why do loan sharks worry less about moral hazard in connection with their borrowers than some other lenders do?
10. If you are an employer, what kinds of moral hazard problems might you worry about with your employees?

- *11. If there were no asymmetry in the information that a borrower and a lender had, could there still be a moral hazard problem?
- 12. "In a world without information and transaction costs, financial intermediaries would not exist." Is this statement true, false, or uncertain? Explain your answer.
- *13. Why might you be willing to make a loan to your neighbour by putting funds in a savings account earning a 5% interest rate at the bank and having the bank lend her the funds at a 10% interest rate rather than lend her the funds yourself?
- 14. How does risk sharing benefit both financial intermediaries and private investors?
- *15. Discuss some of the manifestations of the globalization of world capital markets.

WEB EXERCISES

1. One of the single best sources of information about financial institutions is the financial data produced by the OSFI. Go to **www.osfi-bsif.gc.ca**, click on "Banks" and then "Financial Data-Banks" and answer the following.
 - a. What percent of assets do domestic chartered banks hold in loans? What percentage of assets are held in mortgage loans?
 - b. What percent of assets do trust companies hold in mortgage loans?
 - c. What percent of assets do cooperatives hold in mortgage loans and in consumer loans?
2. The most famous financial market in the world is the New York Stock Exchange. Go to **www.nyse.com**.
 - a. Click on "Information for Media" and summarize the activity in the market in terms of movements in the Dow Jones Industrial Average and the NYSE Composite Index.
 - b. Firms must pay a fee to list their shares for sale on the NYSE. What would be the fee for a firm with 5 million common shares outstanding?



Be sure to visit the MyEconLab website at **www.myeconlab.com**. This online homework and tutorial system puts you in control of your own learning with study and practice tools directly correlated to this chapter content.

CHAPTER 9

An Economic Analysis of Financial Structure

LEARNING OBJECTIVES

After studying this chapter you should be able to

1. depict how asymmetric information results in adverse selection and moral hazard problems that interfere with the efficient functioning of financial markets
2. express how government regulation, the private production and sale of information, and financial intermediaries can lessen, but cannot eliminate, asymmetric information problems
3. discuss why securities regulators are introducing new rules and regulations, such as the Sarbanes-Oxley Act in the United States

PREVIEW

A healthy and vibrant economy requires a financial system that moves funds from people who save to people who have productive investment opportunities. But how does the financial system make sure that your hard-earned savings get channelled to Paula the Productive Investor rather than to Benny the Bum?

This chapter answers that question by providing an economic analysis of how our financial structure is designed to promote economic efficiency. The analysis focuses on a few simple but powerful economic concepts that enable us to explain features of our financial system such as why financial contracts are written as they are and why financial intermediaries are more important than securities markets for getting funds to borrowers. The analysis also demonstrates the important link between the financial system and the performance of the aggregate economy, which is the subject of the last part of the book.

BASIC FACTS ABOUT FINANCIAL STRUCTURE THROUGHOUT THE WORLD

The financial system is complex in structure and function throughout the world. It includes many different types of institutions: banks, insurance companies, mutual funds, stock and bond markets, and so on—all of which are regulated by government. The financial system channels billions of dollars per year from savers to people with productive investment opportunities. If we take a close look at financial structure all over the world, we find eight basic facts, some of which are quite surprising, that we need to explain in order to understand how the financial system works.

The bar chart in Figure 8-1 shows how Canadian businesses financed their activities using external funds (those obtained from outside the business itself) in the period 1970–2002 and compares the Canadian data to those of Germany, Japan, and the United States. The *Bank Loans* category is made up primarily of loans from depository institutions; *Nonbank Loans* is composed primarily of loans by other financial intermediaries; the *Bonds* category includes marketable debt securities such as corporate bonds and commercial paper; and *Stock* consists of new issues of new equity (stock market shares).

Now let us explore the eight facts.

1. ***Stocks are not the most important source of external financing for businesses.*** Because so much attention in the media is focused on the stock market, many people have the impression that stocks are the most important sources of financing for Canadian corporations. However, as we can see from the bar chart in Figure 8-1, the stock market accounted for only a small fraction of the external financing of businesses in the 1970–2002 period: 12%.¹

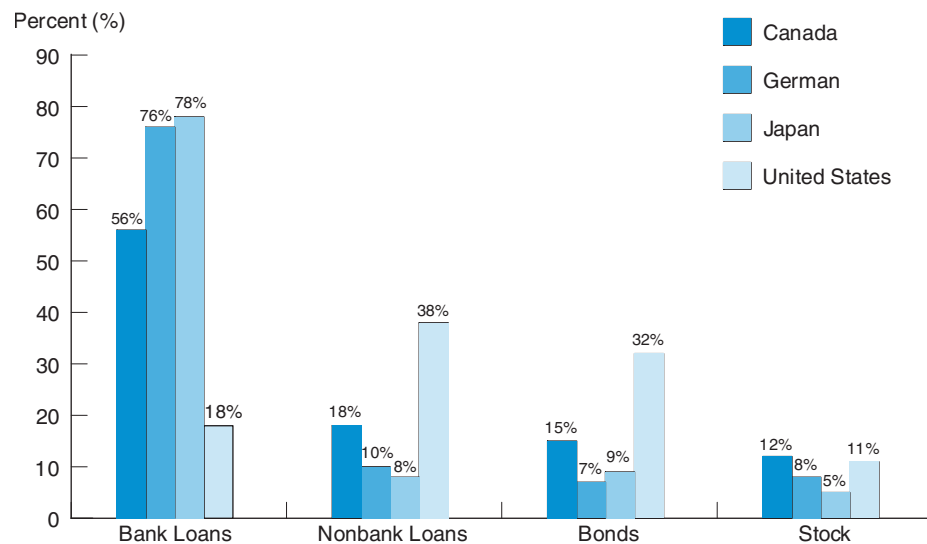


FIG RE 9-1 Sources of External Funds for Nonfinancial Businesses: A Comparison of Canada with Germany, Japan, and the United States

The data are for the 1970–2002 period for Canada and for the 1970–2000 period for Germany, Japan, and the United States.

Sources: Andreas Hackethal and Reinhard H. Schmidt, “Financing Patterns: Measuring Concepts and Empirical Results,” Johann Wolfgang Goethe-Universität Working Paper No. 125, January 2004; and Apostolos Serletis and Karl Pinno, “Corporate Financing in Canada,” *Journal of Economic Asymmetries* 3 (2006): 1–20.

¹ The 12% figure for the percentage of external financing provided by stocks is based on the flows of external funds to corporations. However, this flow figure is somewhat misleading, because when a share of stock is issued, it raises funds permanently, whereas when a bond is issued, it raises funds only temporarily until they are paid back at maturity. To see this, suppose that a firm raises \$1000 by selling a share of stock and another \$1000 by selling a \$1000 one-year bond. In the case of the stock issue, the firm can hold on to the \$1000 it raised this way, but to hold on to the \$1000 it raised through debt, it has to issue a new \$1000 bond every year. If we look at the flow of funds to corporations over a 33-year period, as in Figure 8-1, the firm will have raised \$1000 with a stock issue only once in the 33-year period, while it will have raised \$1000 with debt 33 times, once in each of the 33 years. Thus it will look as though debt is 33 times more important than stocks in raising funds, even though our example indicates that they are actually equally important for the firm.

Similarly small figures apply in the other countries presented in Figure 8-1 as well. Why is the stock market less important than other sources of financing in Canada and other countries?

2. **Issuing marketable debt and equity securities is not the primary way in which businesses finance their operations.** Figure 8-1 shows that bonds are a more important source of financing than stocks in Canada (15% versus 12%). However, stocks and bonds combined (27%), which make up the total share of marketable securities, still supply less than one-third of the external funds corporations need to finance their activities. The fact that issuing marketable securities is not the most important source of financing is true elsewhere in the world as well. Indeed, as we see in Figure 8-1, other countries (except the United States) have a much smaller share of external financing supplied by marketable securities than Canada. Why don't businesses use marketable securities more extensively to finance their activities?
3. **Indirect finance, which involves the activities of financial intermediaries, is many times more important than direct finance, in which businesses raise funds directly from lenders in financial markets.** Direct finance involves the sale to households of marketable securities such as stocks and bonds. The 27% share of stocks and bonds as a source of external financing for Canadian businesses actually greatly overstates the importance of direct finance in our financial system. In general, only a small fraction of newly issued corporate bonds, commercial paper, and stocks are sold directly to Canadian households. The rest of these securities are bought primarily by financial intermediaries such as insurance companies, pension funds, and mutual funds. Because in most countries marketable securities are an even less important source of finance than in Canada, direct finance is also far less important than indirect finance in the rest of the world. Why are financial intermediaries and indirect finance so important in financial markets? In recent years, indirect finance has been declining in importance. Why is this happening?
4. **Financial intermediaries, particularly banks, are the most important source of external funds used to finance businesses.** As we can see in Figure 8-1, the primary source of external funds for businesses throughout the world are loans made by banks and other nonbank financial intermediaries such as insurance companies, pension funds, and finance companies (56% in the United States, but over 70% in Japan, Germany, and Canada). In other industrialized countries, bank loans are the largest category of sources of external finance and so the data suggest that banks in these countries have the most important role in financing business activities. In developing countries, banks play an even more important role in the financial system than they do in the industrialized countries. What makes banks so important to the workings of the financial system? Although banks remain important, their share of external funds for businesses has been declining in recent years. What is driving their decline?
5. **The financial system is among the most heavily regulated sectors of the economy.** The financial system is heavily regulated in Canada and all other developed countries. Governments regulate financial markets primarily to promote the provision of information, and to ensure the soundness (stability) of the financial system. Why are financial markets so extensively regulated throughout the world?

6. ***Only large, well-established corporations have easy access to securities markets to finance their activities.*** Individuals and smaller businesses that are not well established are less likely to raise funds by issuing marketable securities. Instead, they most often obtain their financing from banks. Why do only large, well-known corporations find it easier to raise funds in securities markets?
7. ***Collateral is a prevalent feature of debt contracts for both households and businesses.*** **Collateral** is property that is pledged to a lender to guarantee payment in the event that the borrower is unable to make debt payments. Collateralized debt (also known as **secured debt** to contrast it with **unsecured debt**, such as credit card debt, which is not collateralized) is the predominant form of household debt and is widely used in business borrowing as well. The majority of household debt in Canada consists of collateralized loans: Your automobile is collateral for your auto loan, and your house is collateral for your mortgage. Commercial and farm mortgages, for which property is pledged as collateral, make up one-quarter of borrowing by nonfinancial businesses; corporate bonds and other bank loans also often involve pledges of collateral. Why is collateral such an important feature of debt contracts?
8. ***Debt contracts typically are extremely complicated legal documents that place substantial restrictions on the behaviour of the borrower.*** Many students think of a debt contract as a simple IOU that can be written on a single piece of paper. The reality of debt contracts is far different, however. In all countries, bond or loan contracts typically are long legal documents with provisions (called **restrictive covenants**) that restrict and specify certain activities that the borrower can engage in. Restrictive covenants are not just a feature of debt contracts for businesses; for example, personal automobile loan and home mortgage contracts have covenants that require the borrower to maintain sufficient insurance on the automobile or house purchased with the loan. Why are debt contracts so complex and restrictive?

As you may recall from Chapter 2, an important feature of financial markets is that they have substantial transaction and information costs. An economic analysis of how these costs affect financial markets provides us with explanations of the eight facts, which in turn provide us with a much deeper understanding of how our financial system works. In the next section, we examine the impact of transaction costs on the structure of our financial system. Then we turn to the effect of information costs on financial structure.

TRANSACTION COSTS

Transaction costs are a major problem in financial markets. An example will make this clear.

How Transaction Costs Influence Financial Structure

Say you have \$500 you would like to invest, and you think about investing in the stock market. Because you have only \$500, you can buy only a small number of shares. Even if you use online trading, your purchase is so small that the brokerage commission for buying the stock you picked will be a large percentage of the purchase price of the shares. If instead you decide to buy a bond, the problem is even worse because the smallest denomination for some bonds you might want to buy

is as much as \$10 000 and you do not have that much to invest. You are disappointed and realize that you will not be able to use financial markets to earn a return on your hard-earned savings. You can take some consolation, however, in the fact that you are not alone in being stymied by high transaction costs. This is a fact of life for many of us.

You also face another problem because of transaction costs. Because you have only a small amount of funds available, you can make only a restricted number of investments, because a large number of small transactions would result in very high transaction costs. That is, you have to put all your eggs in one basket, and your inability to diversify will subject you to a lot of risk.

How Financial Intermediaries Reduce Transaction Costs

This example of the problems posed by transaction costs and the example outlined in Chapter 2 when legal costs kept you from making a loan to Carl the Carpenter illustrate that small savers like you are frozen out of financial markets and are unable to benefit from them. Fortunately, financial intermediaries, an important part of the financial structure, have evolved to reduce transaction costs and allow small savers and borrowers to benefit from the existence of financial markets.

ECONOMIES OF SCALE One solution to the problem of high transaction costs is to bundle the funds of many investors together so that they can take advantage of *economies of scale*, the reduction in transaction costs per dollar of investment as the size (scale) of transactions increases. Bundling investors' funds together reduces transaction costs for each individual investor. Economies of scale exist because the total cost of carrying out a transaction in financial markets increases only a little as the size of the transaction grows. For example, the cost of arranging a purchase of 10 000 shares of stock is not much greater than the cost of arranging a purchase of 50 shares of stock.

The presence of economies of scale in financial markets helps explain why financial intermediaries developed and have become such an important part of our financial structure. The clearest example of a financial intermediary that arose because of economies of scale is a mutual fund. A *mutual fund* is a financial intermediary that sells shares to individuals and then invests the proceeds in bonds or stocks. Because it buys large blocks of stocks or bonds, a mutual fund can take advantage of lower transaction costs. These cost savings are then passed on to individual investors after the mutual fund has taken its cut in the form of management fees for administering their accounts. An additional benefit for individual investors is that a mutual fund is large enough to purchase a widely diversified portfolio of securities. The increased diversification for individual investors reduces their risk, thus making them better off.

Economies of scale are also important in lowering the costs of things such as computer technology that financial institutions need to accomplish their tasks. Once a large mutual fund has invested a lot of money in setting up a telecommunications system, for example, the system can be used for a huge number of transactions at a low cost per transaction.

EXPERTISE Financial intermediaries also arise because they are better able to develop expertise to lower transaction costs. Their expertise in computer technology enables them to offer customers convenient services like being able to call a toll-free number for information on how well their investments are doing and to write cheques on their accounts.

An important outcome of a financial intermediary's low transaction costs is the ability to provide its customers with *liquidity services*, services that make it easier for customers to conduct transactions. Some money market mutual funds, for example, not only pay shareholders high interest rates, but also allow them to write cheques for convenient bill-paying.

ASYMMETRIC INFORMATION: ADVERSE SELECTION AND MORAL HAZARD

The presence of transaction costs in financial markets explains in part why financial intermediaries and indirect finance play such an important role in financial markets (fact 3). To understand financial structure more fully, however, we turn to the role of information in financial markets.²

Asymmetric information—one party's having insufficient knowledge about the other party involved in a transaction to make accurate decisions—is an important aspect of financial markets. For example, managers of a corporation know whether they are honest or have better information about how well their business is doing than the stockholders do. The presence of asymmetric information leads to adverse selection and moral hazard problems, which were introduced in Chapter 2.

Adverse selection is an asymmetric information problem that occurs *before* the transaction occurs: potential bad credit risks are the ones who most actively seek out loans. Thus the parties who are the most likely to produce an undesirable outcome are the ones most likely to want to engage in the transaction. For example, big risk takers or outright crooks might be the most eager to take out a loan because they know that they are unlikely to pay it back. Because adverse selection increases the chances that a loan might be made to a bad credit risk, lenders may decide not to make any loans even though there are good credit risks in the marketplace.

Moral hazard arises *after* the transaction occurs: the lender runs the risk that the borrower will engage in activities that are undesirable from the lender's point of view because they make it less likely that the loan will be paid back. For example, once borrowers have obtained a loan, they may take on big risks (which have possible high returns but also run a greater risk of default) because they are playing with someone else's money. Because moral hazard lowers the probability that the loan will be repaid, lenders may decide that they would rather not make a loan.

The analysis of how asymmetric information problems affect economic behaviour is called **agency theory**. We will apply this theory here to explain why financial structure takes the form it does, thereby explaining the facts described at the beginning of the chapter.

THE LEMONS PROBLEM: HOW ADVERSE SELECTION INFLUENCES FINANCIAL STRUCTURE

A particular characterization of how the adverse selection problem interferes with the efficient functioning of a market was outlined in a famous article by Nobel Prize winner George Akerlof. It is called the “lemons problem” because it resem-

² An excellent survey of the literature on information and financial structure that expands on the topics discussed in the rest of this chapter is contained in Mark Gertler, “Financial Structure and Aggregate Economic Activity: An Overview,” *Journal of Money, Credit and Banking* 20 (1988): 559–588.

bles the problem created by lemons in the used-car market.³ Potential buyers of used cars are frequently unable to assess the quality of the car; that is, they can't tell whether a particular used car is a car that will run well or a lemon that will continually give them grief. The price that a buyer pays must therefore reflect the *average* quality of the cars in the market, somewhere between the low value of a lemon and the high value of a good car.

The owner of a used car, by contrast, is more likely to know whether the car is a peach or a lemon. If the car is a lemon, the owner is more than happy to sell it at the price the buyer is willing to pay, which, being somewhere between the value of a lemon and a good car, is greater than the lemon's value. However, if the car is a peach, the owner knows that the car is undervalued at the price the buyer is willing to pay, and so the owner may not want to sell it. As a result of this adverse selection, few good used cars will come to the market. Because the average quality of a used car available in the market will be low and because few people want to buy a lemon, there will be few sales. The used-car market will function poorly, if at all.

Lemons in the Stock and Bond Markets

A similar lemons problem arises in securities markets, that is, the debt (bond) and equity (stock) markets. Suppose that our friend Irving the Investor, a potential buyer of securities such as common stock, can't distinguish between good firms with high expected profits and low risk and bad firms with low expected profits and high risk. In this situation, Irving will be willing to pay only a price that reflects the *average* quality of firms issuing securities—a price that lies between the value of securities from bad firms and the value of those from good firms. If the owners or managers of a good firm have better information than Irving and *know* that they are a good firm, they know that their securities are undervalued and will not want to sell them to Irving at the price he is willing to pay. The only firms willing to sell Irving securities will be bad firms (because his price is higher than the securities are worth). Our friend Irving is not stupid; he does not want to hold securities in bad firms, and hence he will decide not to purchase securities in the market. In an outcome similar to that in the used-car market, this securities market will not work very well because few firms will sell securities in it to raise capital.

The analysis is similar if Irving considers purchasing a corporate debt instrument in the bond market rather than an equity share. Irving will buy a bond only if its interest rate is high enough to compensate him for the average default risk of the good and bad firms trying to sell the debt. The knowledgeable owners of a good firm realize that they will be paying a higher interest rate than they should, and so they are unlikely to want to borrow in this market. Only the bad firms will be willing to borrow, and because investors like Irving are not eager to buy bonds issued by bad firms, they will probably not buy any bonds at all. Few bonds are likely to sell in this market, and so it will not be a good source of financing.

³ George Akerlof, "The Market for 'Lemons': Quality, Uncertainty and the Market Mechanism," *Quarterly Journal of Economics* 84 (1970): 488–500. Two important papers that have applied the lemons problem analysis to financial markets are Stewart Myers and N. S. Majluf, "Corporate Financing and Investment Decisions When Firms Have Information That Investors Do Not Have," *Journal of Financial Economics* 13 (1984): 187–221, and Bruce Greenwald, Joseph E. Stiglitz, and Andrew Weiss, "Information Imperfections in the Capital Market and Macroeconomic Fluctuations," *American Economic Review* 74 (1984): 194–199.

The analysis we have just conducted explains fact 2—why marketable securities are not the primary source of financing for businesses in any country in the world. It also partly explains fact 1—why stocks are not the most important source of financing for Canadian businesses. The presence of the lemons problem keeps securities markets such as the stock and bond markets from being effective in channelling funds from savers to borrowers.

Tools to Help Solve Adverse Selection Problems

In the absence of asymmetric information, the lemons problem goes away. If buyers know as much about the quality of used cars as sellers so that all involved can tell a good car from a bad one, buyers will be willing to pay full value for good used cars. Because the owners of good used cars can now get a fair price, they will be willing to sell them in the market. The market will have many transactions and will do its intended job of channelling good cars to people who want them.

Similarly, if purchasers of securities can distinguish good firms from bad, they will pay the full value of securities issued by good firms, and good firms will sell their securities in the market. The securities market will then be able to move funds to the good firms that have the most productive investment opportunities.

PRIVATE PRODUCTION AND SALE OF INFORMATION The solution to the adverse selection problem in financial markets is to eliminate asymmetric information by furnishing people supplying funds with full details about the individuals or firms seeking to finance their investment activities. One way to get this material to saver-lenders is to have private companies collect and produce information that distinguishes good from bad firms and then sell it. In Canada, companies such as Standard & Poor's and the Dominion Bond Rating Service gather information on firms' balance sheet positions and investment activities, publish these data, and sell them to subscribers (individuals, libraries, and financial intermediaries involved in purchasing securities).

The system of private production and sale of information does not completely solve the adverse selection problem in securities markets, however, because of the so-called **free-rider problem**. The free-rider problem occurs when people who do not pay for information take advantage of the information that other people have paid for. The free-rider problem suggests that the private sale of information will be only a partial solution to the lemons problem. To see why, suppose that you have just purchased information that tells you which firms are good and which are bad. You believe that this purchase is worthwhile because you can make up the cost of acquiring this information, and then some, by purchasing the securities of good firms that are undervalued. However, when our savvy (free-riding) investor Irving sees you buying certain securities, he buys right along with you, even though he has not paid for any information. If many other investors act as Irving does, the increased demand for the undervalued good securities will cause their low price to be bid up immediately to reflect the securities' true value. Because of all these free riders, you can no longer buy the securities for less than their true value. Now because you will not gain any profits from purchasing the information, you realize that you never should have paid for this information in the first place. If other investors come to the same realization, private firms and individuals may not be able to sell enough of this information to make it worth their while to gather and produce it. The weakened ability of private firms to profit from selling information will mean that less information is produced in the marketplace, and so adverse selection (the lemons problem) will still interfere with the efficient functioning of securities markets.

GOVERNMENT REGULATION TO INCREASE INFORMATION The free-rider problem prevents the private market from producing enough information to eliminate all the asymmetric information that leads to adverse selection. Could financial markets benefit from government intervention? The government could, for instance, produce information to help investors distinguish good from bad firms and provide it to the public free of charge. This solution, however, would involve the government in releasing negative information about firms, a practice that might be politically difficult. A second possibility (and one followed by Canada and most governments throughout the world) is for the government to regulate securities markets in a way that encourages firms to reveal honest information about themselves so that investors can determine how good or bad the firms are. In Canada, government regulation exists that requires firms selling securities to have independent **a dits**, in which accounting firms certify that the firm adheres to standard accounting principles and discloses information about sales, assets, and earnings. Similar regulations are found in other countries. However, disclosure requirements do not always work well, as the recent collapse of Enron and accounting scandals at other corporations, such as WorldCom and Parmalat (an Italian company) suggest (see the FYI box, The Enron Implosion).

The asymmetric information problem of adverse selection in financial markets helps explain why financial markets are among the most heavily regulated sectors in the economy (fact 5). Government regulation to increase information for investors is needed to reduce the adverse selection problem, which interferes with the efficient functioning of securities (stock and bond) markets.

FYI

The Enron Implosion

Until 2001, Enron Corporation, a firm that specialized in trading in the energy market, appeared to be spectacularly successful. It had a quarter of the energy-trading market and was valued as high as US\$77 billion in August 2000 (just a little over a year before its collapse), making it the seventh largest corporation in the United States at that time. Toward the end of 2001, however, Enron came crashing down. In October 2001, Enron announced a third-quarter loss of US\$618 million and disclosed accounting “mistakes.” The U.S. SEC then engaged in a formal investigation of Enron’s financial dealings with partnerships led by its former finance chief. It became clear that Enron was engaged in a complex set of transactions by which it was keeping substantial amounts of debt and financial contracts off its balance sheet. These transactions enabled Enron to hide its financial difficulties.

Despite securing as much as US\$1.5 billion of new financing from JPMorgan Chase and Citigroup, the company was forced to declare bankruptcy in December 2001, making it the largest bankruptcy in U.S. history.

The Enron collapse illustrates that government regulation can lessen asymmetric information problems but cannot eliminate them. Managers have tremendous incentives to hide their companies’ problems, making it hard for investors to know the true value of the firm.

The Enron bankruptcy not only increased concerns in financial markets about the quality of accounting information supplied by corporations, but it also led to hardship for many of the former employees who found that their pensions had become worthless. Outrage against executives at Enron was high, and several were indicted, convicted, and sent to jail.

Although government regulation lessens the adverse selection problem, it does not eliminate it. Even when firms provide information to the public about their sales, assets, or earnings, they still have more information than investors: there is a lot more to knowing the quality of a firm than statistics can provide. Furthermore, bad firms have an incentive to make themselves look like good firms because this would enable them to fetch a higher price for their securities. Bad firms will slant the information they are required to transmit to the public, thus making it harder for investors to sort out the good firms from the bad.

FINANCIAL INTERMEDIATION So far we have seen that private production of information and government regulation to encourage provision of information lessen but do not eliminate the adverse selection problem in financial markets. How, then, can the financial structure help promote the flow of funds to people with productive investment opportunities when there is asymmetric information? A clue is provided by the structure of the used-car market.

An important feature of the used-car market is that most used cars are not sold directly by one individual to another. An individual considering buying a used car might pay for privately produced information by subscribing to a magazine like *Consumer Reports* to find out if a particular make of car has a good repair record. Nevertheless, reading *Consumer Reports* does not solve the adverse selection problem because even if a particular make of car has a good reputation, the specific car someone is trying to sell could be a lemon. The prospective buyer might also bring the used car to a mechanic for a once-over. But what if the prospective buyer doesn't know a mechanic who can be trusted or if the mechanic charges a high fee to evaluate the car?

Because these roadblocks make it hard for individuals to acquire enough information about used cars, most used cars are not sold directly by one individual to another. Instead, they are sold by an intermediary, a used-car dealer who purchases used cars from individuals and resells them to other individuals. Used-car dealers produce information in the market by becoming experts in determining whether a car is a peach or a lemon. Once they know that a car is good, they can sell it with some form of a guarantee: either a guarantee that is explicit, such as a warranty, or an implicit guarantee in which they stand by their reputation for honesty. People are more likely to purchase a used car because of a dealer's guarantee, and the dealer is able to make a profit on the production of information about automobile quality by being able to sell the used car at a higher price than the dealer paid for it. If dealers purchase and then resell cars on which they have produced information, they avoid the problem of other people free-riding on the information they produced.

Just as used-car dealers help solve adverse selection problems in the automobile market, financial intermediaries play a similar role in financial markets. A financial intermediary such as a bank becomes an expert in producing information about firms so that it can sort out good credit risks from bad ones. Then it can acquire funds from depositors and lend them to the good firms. Because the bank is able to lend mostly to good firms, it is able to earn a higher return on its loans than the interest it has to pay to its depositors. The resulting profit that the bank earns allows it to engage in this information production activity.

An important element in the bank's ability to profit from the information it produces is that it avoids the free-rider problem by primarily making private loans rather than by purchasing securities that are traded in the open market. Because a private loan is not traded, other investors cannot watch what the bank is doing

and bid up the loan's price to the point that the bank receives no compensation for the information it has produced. The bank's role as an intermediary that holds mostly nontraded loans is the key to its success in reducing asymmetric information in financial markets.

Our analysis of adverse selection indicates that financial intermediaries in general, and banks in particular because they hold a large fraction of nontraded loans, should play a greater role in moving funds to corporations than securities markets do. Our analysis thus explains facts 3 and 4: why indirect finance is so much more important than direct finance and why banks are the most important source of external funds for financing businesses.

Another important fact that is explained by the analysis here is the greater importance of banks in the financial systems of developing countries. As we have seen, when the quality of information about firms is better, asymmetric information problems will be less severe, and it will be easier for firms to issue securities. Information about private firms is harder to collect in developing countries than in industrialized countries; therefore, the smaller role played by securities markets leaves a greater role for financial intermediaries such as banks. A corollary of this analysis is that as information about firms becomes easier to acquire, the role of banks should decline. A major development in the past 20 years has been huge improvements in information technology. Thus the analysis here suggests that the lending role of financial institutions such as banks should have declined, and this is exactly what has occurred.

Our analysis of adverse selection also explains fact 6, which questions why large firms are more likely to obtain funds from securities markets, a direct route, rather than from banks and financial intermediaries, an indirect route. The better known a corporation is, the more information about its activities is available in the marketplace. Thus it is easier for investors to evaluate the quality of the corporation and determine whether it is a good firm or a bad one. Because investors have fewer worries about adverse selection with well-known corporations, they will be willing to invest directly in their securities. Our adverse selection analysis thus suggests that there should be a pecking order for firms that can issue securities. Hence we have an explanation for fact 6: The larger and more established a corporation is, the more likely it will be to issue securities to raise funds.

COLLATERAL AND NET WORTH Adverse selection interferes with the functioning of financial markets only if a lender suffers a loss when a borrower is unable to make loan payments and thereby defaults. Collateral, property promised to the lender if the borrower defaults, reduces the consequences of adverse selection because it reduces the lender's losses in the event of a default. If a borrower defaults on a loan, the lender can sell the collateral and use the proceeds to make up for the losses on the loan. For example, if you fail to make your mortgage payments, the lender can take title to your house, auction it off, and use the receipts to pay off the loan. Lenders are thus more willing to make loans secured by collateral, and borrowers are willing to supply collateral because the reduced risk for the lender makes it more likely they will get the loan in the first place and perhaps at a better loan rate. The presence of adverse selection in credit markets thus provides an explanation for why collateral is an important feature of debt contracts (fact 7).

Net worth (also called **equity capital**), the difference between a firm's assets (what it owns or is owed) and its liabilities (what it owes), can perform a similar role to collateral. If a firm has a high net worth, then even if it engages in invest-

ments that cause it to have negative profits and so defaults on its debt payments, the lender can take title to the firm's net worth, sell it off, and use the proceeds to recoup some of the losses from the loan. In addition, the more net worth a firm has in the first place, the less likely it is to default because the firm has a cushion of assets that it can use to pay off its loans. Hence when firms seeking credit have high net worth, the consequences of adverse selection are less important and lenders are more willing to make loans. This analysis lies behind the often-heard lament, "Only the people who don't need money can borrow it!"

Summary

So far we have used the concept of adverse selection to explain seven of the eight facts about financial structure introduced earlier: the first four emphasize the importance of financial intermediaries and the relative unimportance of securities markets for the financing of corporations; the fifth, that financial markets are among the most heavily regulated sectors of the economy; the sixth, that only large, well-established corporations have access to securities markets; and the seventh, that collateral is an important feature of debt contracts. In the next section we will see that the other asymmetric information concept of moral hazard provides additional reasons for the importance of financial intermediaries and the relative unimportance of securities markets for the financing of corporations, the prevalence of government regulation, and the importance of collateral in debt contracts. In addition, the concept of moral hazard can be used to explain our final fact (fact 8) of why debt contracts are complicated legal documents that place substantial restrictions on the behaviour of borrowers.

HOW MORAL HAZARD AFFECTS THE CHOICE BETWEEN DEBT AND EQUITY CONTRACTS

Moral hazard is the asymmetric information problem that occurs after the financial transaction takes place, when the seller of a security may have incentives to hide information and engage in activities that are undesirable for the purchaser of the security. Moral hazard has important consequences for whether a firm finds it easier to raise funds with debt than with equity contracts.

Moral Hazard in Equity Contracts: The Principal-Agent Problem

Equity contracts, such as common stock, are claims to a share in the profits and assets of a business. Equity contracts are subject to a particular type of moral hazard called the **principal-agent problem**. When managers own only a small fraction of the firm they work for, the stockholders who own most of the firm's equity (called the *principals*) are not the same people as the managers of the firm, who are the *agents* of the owners. This separation of ownership and control involves moral hazard in that the managers in control (the agents) may act in their own interest rather than in the interest of the stockholder-owners (the principals) because the managers have less incentive to maximize profits than the stockholder-owners do.

To understand the principal-agent problem more fully, suppose that your friend Steve asks you to become a silent partner in his ice-cream store. The store requires an investment of \$10 000 to set up and Steve has only \$1000. So you purchase an equity stake (shares) for \$9000, which entitles you to 90% of the ownership of the firm, while Steve owns only 10%. If Steve works hard to make tasty ice cream, keeps the store clean, smiles at all the customers, and hustles to wait on tables quickly, after all expenses (including Steve's salary), the store will have

\$50 000 in profits per year, of which Steve receives 10% (\$5000) and you receive 90% (\$45 000).

But if Steve doesn't provide quick and friendly service to his customers, uses the \$50 000 in income to buy artwork for his office, and even sneaks off to the beach while he should be at the store, the store will not earn any profit. Steve can earn the additional \$5000 (his 10% share of the profits) over his salary only if he works hard and forgoes unproductive investments (such as art for his office). Steve might decide that the extra \$5000 just isn't enough to make him want to expend the effort to be a good manager; he might decide that it would be worth his while only if he earned an extra \$10 000. If Steve feels this way, he does not have enough incentive to be a good manager and will end up with a beautiful office, a good tan, and a store that doesn't show any profits. Because the store won't show any profits, Steve's decision not to act in your interest will cost you \$45 000 (your 90% of the profits if he had chosen to be a good manager instead).

The moral hazard arising from the principal-agent problem might be even worse if Steve were not totally honest. Because his ice-cream store is a cash business, Steve has the incentive to pocket \$50 000 in cash and tell you that the profits were zero. He now gets a return of \$50 000, and you get nothing.

Further indications that the principal-agent problem created by equity contracts can be severe are provided by recent corporate scandals in corporations such as Enron and Tyco International, in which managers have been accused of diverting funds for personal use. Besides pursuing personal benefits, managers might also pursue corporate strategies (such as the acquisition of other firms) that enhance their personal power but do not increase the corporation's profitability.

The principal-agent problem would not arise if the owners of a firm had complete information about what the managers were up to and could prevent wasteful expenditures or fraud. The principal-agent problem, which is an example of moral hazard, arises only because a manager, like Steve, has more information about his activities than the stockholder does—that is, there is asymmetric information. The principal-agent problem would also not arise if Steve alone owned the store and there were no separation of ownership and control. If this were the case, Steve's hard work and avoidance of unproductive investments would yield him a profit (and extra income) of \$50 000, an amount that would make it worth his while to be a good manager.

Tools to Help Solve the Principal– Agent Problem

PRODUCTION OF INFORMATION: MONITORING You have seen that the principal-agent problem arises because managers have more information about their activities and actual profits than stockholders do. One way for stockholders to reduce this moral hazard problem is for them to engage in a particular type of information production, the monitoring of the firm's activities: auditing the firm frequently and checking on what the management is doing. The problem is that the monitoring process can be expensive in terms of time and money, as reflected in the name economists give it, **costly state verification**. Costly state verification makes the equity contract less desirable, and it explains, in part, why equity is not a more important element in our financial structure.

As with adverse selection, the free-rider problem decreases the amount of information production that would reduce the moral hazard (principal-agent) problem. In this example, the free-rider problem decreases monitoring. If you know that other stockholders are paying to monitor the activities of the company you hold shares in, you can take a free ride on their activities. Then you can use the money you save by not engaging in monitoring to vacation on a Caribbean

island. If you can do this, though, so can other stockholders. Perhaps all the stockholders will go to the islands, and no one will spend any resources on monitoring the firm. The moral hazard problem for shares of common stock will then be severe, making it hard for firms to issue them to raise capital (providing an explanation for fact 1).

GOVERNMENT REGULATION TO INCREASE INFORMATION As with adverse selection, the government has an incentive to try to reduce the moral hazard problem created by asymmetric information, which provides another reason why the financial system is so heavily regulated (fact 5). Governments everywhere have laws to force firms to adhere to standard accounting principles that make profit verification easier. They also pass laws to impose stiff criminal penalties on people who commit the fraud of hiding and stealing profits. However, these measures can only be partly effective. Catching this kind of fraud is not easy; fraudulent managers have the incentive to make it very hard for government agencies to find or prove fraud.

FINANCIAL INTERMEDIATION Financial intermediaries have the ability to avoid the free-rider problem in the face of moral hazard, and this is another reason why indirect finance is so important (fact 3). One financial intermediary that helps reduce the moral hazard arising from the principal–agent problem is the **venture capital firm**. Venture capital firms pool the resources of their partners and use the funds to help budding entrepreneurs start new businesses. In exchange for the use of the venture capital, the firm receives an equity share in the new business. Because verification of earnings and profits is so important in eliminating moral hazard, venture capital firms usually insist on having several of their own people participate as members of the managing body of the firm, the board of directors, so that they can keep a close watch on the firm's activities. When a venture capital firm supplies start-up funds, the equity in the firm is not marketable to anyone *but* the venture capital firm. Thus other investors are unable to take a free ride on the venture capital firm's verification activities. As a result of this arrangement, the venture capital firm is able to garner the full benefits of its verification activities and is given the appropriate incentives to reduce the moral hazard problem.

Venture capital firms have been important in the development of the high-tech sector in Canada and the United States, which has resulted in job creation, economic growth, and increased international competitiveness.

DEBT CONTRACTS Moral hazard arises with an equity contract, which is a claim on profits in all situations, whether the firm is making or losing money. If a contract could be structured so that moral hazard would exist only in certain situations, there would be a reduced need to monitor managers, and the contract would be more attractive than the equity contract. The debt contract has exactly these attributes because it is a contractual agreement by the borrower to pay the lender *fixed* dollar amounts at periodic intervals. When the firm has high profits, the lender receives the contractual payments and does not need to know the exact profits of the firm. If the managers are hiding profits or are pursuing activities that are personally beneficial but don't increase profitability, the lender doesn't care as long as these activities do not interfere with the ability of the firm to make its debt payments on time. Only when the firm cannot meet its debt payments, thereby being in a state of default, is there a need for the lender to verify the state of the

firm's profits. Only in this situation do lenders involved in debt contracts need to act more like equity holders; now they need to know how much income the firm has in order to get their fair share.

The less frequent need to monitor the firm, and thus a lower cost of state verification, helps explain why debt contracts are used more frequently than equity contracts to raise capital. The concept of moral hazard thus helps explain fact 1, why stocks are not the most important source of financing for businesses.⁴

HOW MORAL HAZARD INFLUENCES FINANCIAL STRUCTURE IN DEBT MARKETS

Even with the advantages just described, debt contracts are still subject to moral hazard. Because a debt contract requires the borrowers to pay out a fixed amount and lets them keep any profits above this amount, the borrowers have an incentive to take on investment projects that are riskier than the lenders would like.

For example, suppose that because you are concerned about the problem of verifying the profits of Steve's ice-cream store, you decide not to become an equity partner. Instead, you lend Steve the \$9000 he needs to set up his business and have a debt contract that pays you an interest rate of 10%. As far as you are concerned, this is a surefire investment because there is a strong and steady demand for ice cream in your neighbourhood. However, once you give Steve the funds, he might use them for purposes other than you intended. Instead of opening up the ice-cream store, Steve might use your \$9000 loan to invest in chemical research equipment because he thinks he has a 1-in-10 chance of inventing a diet ice cream that tastes every bit as good as the premium brands but has no fat or calories.

Obviously, this is a very risky investment, but if Steve is successful, he will become a multimillionaire. He has a strong incentive to undertake the riskier investment with your money because the gains to him would be so large if he succeeded. You would clearly be very unhappy if Steve used your loan for the riskier investment because if he were unsuccessful, which is highly likely, you would lose most, if not all, of the money you loaned him. And if he were successful, you wouldn't share in his success—you would still get only a 10% return on the loan because the principal and interest payments are fixed. Because of the potential moral hazard (that Steve might use your money to finance a very risky venture), you would probably not make the loan to Steve, even though an ice-cream store in the neighbourhood is a good investment that would provide benefits for everyone.

Tools to Help Solve Moral Hazard in Debt Contracts

NET WORTH AND COLLATERAL When borrowers have more at stake because their net worth (the difference between their assets and liabilities) or the collateral they have pledged to the lender is high, the risk of moral hazard—the temptation to act in a manner that lenders find objectionable—will be greatly reduced because the borrowers themselves have a lot to lose. Let's return to Steve and his ice-cream business. Suppose that the cost of setting up either the ice-cream store or the

⁴ Another factor that encourages the use of debt contracts rather than equity contracts is our tax laws. Debt interest payments are a deductible expense for Canadian firms, whereas dividend payments to equity shareholders are not.

research equipment is \$100 000 instead of \$10 000. So Steve needs to put \$91 000 of his own money into the business (instead of \$1000) in addition to the \$9000 supplied by your loan. Now if Steve is unsuccessful in inventing the no-calorie nonfat ice cream, he has a lot to lose, the \$91 000 of net worth (\$100 000 in assets minus the \$9000 loan from you). He will think twice about undertaking the riskier investment and is more likely to invest in the ice-cream store, which is more of a sure thing. Hence when Steve has more of his own money (net worth) in the business, you are more likely to make him the loan.

Similarly, if you have pledged your house as collateral, you are less likely to go to Las Vegas and gamble away your earnings that month because you might not be able to make your mortgage payments and might lose your house.

One way of describing the solution that high net worth and collateral provides to the moral hazard problem is to say that it makes the debt contract **incentive-compatible**; that is, it aligns the incentives of the borrower with those of the lender. The greater the borrower's net worth and collateral pledged, the greater the borrower's incentive to behave in the way that the lender expects and desires, the smaller the moral hazard problem in the debt contract and the easier it is for the firm or household to borrow. Conversely, when the borrower's net worth and collateral is lower, the moral hazard problem is greater, and it is harder to borrow.

MONITORING AND ENFORCEMENT OF RESTRICTIVE COVENANTS As the example of Steve and his ice-cream store shows, if you could make sure that Steve doesn't invest in anything riskier than the ice-cream store, it would be worth your while to make him the loan. You can ensure that Steve uses your money for the purpose *you* want it to be used for by writing provisions (restrictive covenants) into the debt contract that restrict his firm's activities. By monitoring Steve's activities to see whether he is complying with the restrictive covenants and enforcing the covenants if he is not, you can make sure that he will not take on risks at your expense. Restrictive covenants are directed at reducing moral hazard either by ruling out undesirable behaviour or by encouraging desirable behaviour. There are four types of restrictive covenants that achieve this objective:

1. ***Co covenants to discourage undesirable behaviour.*** Covenants can be designed to lower moral hazard by keeping the borrower from engaging in the undesirable behaviour of undertaking risky investment projects. Some covenants mandate that a loan can be used only to finance specific activities, such as the purchase of particular equipment or inventories. Others restrict the borrowing firm from engaging in certain risky business activities, such as purchasing other businesses.
2. ***Co covenants to encourage desirable behaviour.*** Restrictive covenants can encourage the borrower to engage in desirable activities that make it more likely that the loan will be paid off. One restrictive covenant of this type requires the breadwinner in a household to carry life insurance that pays off the mortgage upon that person's death. Restrictive covenants of this type for businesses focus on encouraging the borrowing firm to keep its net worth high because higher borrower net worth reduces moral hazard and makes it less likely that the lender will suffer losses. These restrictive covenants typically specify that the firm must maintain minimum holdings of certain assets relative to the firm's size.

3. ***Covenants to keep collateral valuable.*** Because collateral is an important protection for the lender, restrictive covenants can encourage the borrower to keep the collateral in good condition and make sure that it stays in the possession of the borrower. This is the type of covenant ordinary people encounter most often. Automobile loan contracts, for example, require the car owner to maintain a minimum amount of collision and theft insurance and prevent the sale of the car unless the loan is paid off. Similarly, the recipient of a home mortgage must have adequate insurance on the home and must pay off the mortgage when the property is sold.
4. ***Covenants to provide information.*** Restrictive covenants also require a borrowing firm to provide information about its activities periodically in the form of quarterly financial statements, thereby making it easier for the lender to monitor the firm and reduce moral hazard. This type of covenant may also stipulate that the lender has the right to audit and inspect the firm's books at any time.

We now see why debt contracts are often complicated legal documents with numerous restrictions on the borrower's behaviour (fact 8): debt contracts require complicated restrictive covenants to lower moral hazard.

FINANCIAL INTERMEDIATION Although restrictive covenants help reduce the moral hazard problem, they do not eliminate it completely. It is almost impossible to write covenants that rule out *every* risky activity. Furthermore, borrowers may be clever enough to find loopholes in restrictive covenants that make them ineffective.

Another problem with restrictive covenants is that they must be monitored and enforced. A restrictive covenant is meaningless if the borrower can violate it knowing that the lender won't check up or is unwilling to pay for legal recourse. Because monitoring and enforcement of restrictive covenants are costly, the free-rider problem arises in the debt securities (bond) market just as it does in the stock market. If you know that other bondholders are monitoring and enforcing the restrictive covenants, you can free-ride on their monitoring and enforcement. But other bondholders can do the same thing, so the likely outcome is that not enough resources are devoted to monitoring and enforcing the restrictive covenants. Moral hazard therefore continues to be a severe problem for marketable debt.

As we have seen before, financial intermediaries, particularly banks, have the ability to avoid the free-rider problem as long as they primarily make private loans. Private loans are not traded, so no one else can free-ride on the intermediary's monitoring and enforcement of the restrictive covenants. The intermediary making private loans thus receives the benefits of monitoring and enforcement and will work to shrink the moral hazard problem inherent in debt contracts. The concept of moral hazard has provided us with additional reasons why financial intermediaries play a more important role in channelling funds from savers to borrowers than marketable securities do, as described in facts 3 and 4.

Summary

The presence of asymmetric information in financial markets leads to adverse selection and moral hazard problems that interfere with the efficient functioning of those markets. Tools to help solve these problems involve the private production and sale of information, government regulation to increase information in financial markets, the importance of collateral and net worth to debt contracts, and

the use of monitoring and restrictive covenants. A key finding from our analysis is that the existence of the free-rider problem for traded securities such as stocks and bonds indicates that financial intermediaries, particularly banks, should play a greater role than securities markets in financing the activities of businesses. Economic analysis of the consequences of adverse selection and moral hazard has helped explain the basic features of our financial system and has provided solutions to the eight facts about our financial structure outlined at the beginning of this chapter.

To help you keep track of the tools that help solve asymmetric information problems, Table 8-1 provides a listing of the asymmetric information problems and tools that can help solve them. In addition, it lists how these tools and asymmetric information problems explain the eight facts of financial structure described at the beginning of the chapter.

TABLE 9-1 Asymmetric Information Problems and Tools to Solve Them

Asymmetric Information Problem	Tools to Solve It	Explains Fact No.
Adverse selection	Private production and sale of information	1, 2
	Government regulation to increase information	5
	Financial intermediation	3, 4, 6
	Collateral and net worth	7
Moral hazard in equity contracts (principal-agent problem)	Production of information: monitoring	1
	Government regulation to increase information	5
	Financial intermediation	3
	Debt contracts	1
Moral hazard in debt contracts	Net worth and collateral	7
	Monitoring and enforcement of restrictive covenants	8
	Financial intermediation	3, 4

Note: List of facts:

1. Stocks are not the most important source of external financing.
2. Marketable securities are not the primary source of finance.
3. Indirect finance is more important than direct finance.
4. Banks are the most important source of external funds.
5. The financial system is heavily regulated.
6. Only large, well-established firms have easy access to securities markets.
7. Collateral is prevalent in debt contracts.
8. Debt contracts have numerous restrictive covenants.

APPLICATION

Financial Development and Economic Growth

Recent research has found that an important reason why many developing countries or ex-communist countries like Russia (which are referred to as transition countries) experience very low rates of growth is that their financial systems are underdeveloped (a situation referred to as *financial repression*).⁵ The economic analysis of financial structure helps explain how an underdeveloped financial system leads to a low state of economic development and economic growth.

The financial systems in developing and transition countries face several difficulties that keep them from operating efficiently. As we have seen, two important tools used to help solve adverse selection and moral hazard problems in credit markets are collateral and restrictive covenants. In many developing countries, the system of property rights (the rule of law, constraints on government expropriation, absence of corruption) functions poorly, making it hard to make effective use of these two tools. In these countries, bankruptcy procedures are often extremely slow and cumbersome. For example, in many countries, creditors (holders of debt) must first sue the defaulting debtor for payment, which can take several years, and then once a favourable judgement has been obtained, the creditor has to sue again to obtain title to the collateral. The process can take in excess of five years, and by the time the lender acquires the collateral, it may have been neglected and thus have little value. In addition, governments often block lenders from foreclosing on borrowers in politically powerful sectors such as agriculture. Where the market is unable to use collateral effectively, the adverse selection problem will be worse because the lender will need even more information about the quality of the borrower in order to screen out a good loan from a bad one. The result is that it will be harder for lenders to channel funds to borrowers with the most productive investment opportunities. There will be less productive investment and hence a slower-growing economy. Similarly, a poorly developed or corrupt legal system may make it extremely difficult for lenders to enforce restrictive covenants. Thus they may have a much more limited ability to reduce moral hazard on the part of borrowers and so will be less willing to lend. Again the outcome will be less-productive investment and a lower growth rate for the economy. The importance of an effective legal system in promoting economic growth suggests that lawyers play a more positive role in the economy than we give them credit for (see the FYI box, Let the Lawyers Live!).

Governments in developing and transition countries often use their financial systems to direct credit to themselves or to favoured sectors of the economy by setting interest rates at artificially low levels for certain types of loans, by creating so-called development finance institutions to make specific types of loans, or by directing existing institutions to lend to certain entities. As we have seen, private institutions have an incentive to solve adverse selection and moral hazard problems and lend to borrowers with the most productive investment opportunities. Governments have less incentive to do so because they are not driven by the profit motive and so their directed credit programs may not channel funds to sectors that will produce high growth for the economy. The outcome is again likely to result in less-efficient investment and slower growth.

⁵ See World Bank, *Finance for Growth: Policy Choices in a Volatile World* (World Bank and Oxford University Press, 2001) for a survey of the literature linking economic growth to financial development and a list of additional references.

In addition, banks in many developing and transition countries are owned by their governments. Again because of the absence of the profit motive, these **state-owned banks** have little incentive to allocate their capital to the most productive uses. Not surprisingly, the primary loan customer of these state-owned banks is often the government, which does not always use the funds wisely.

FYI

Let the Lawyers Live!

Lawyers are often an easy target for would-be comedians. Countless jokes centre on ambulance-chasing and shifty filers of frivolous lawsuits. Hostility to lawyers is not just a recent phenomenon: in Shakespeare's *Henry VI*, written in the late sixteenth century, Dick the Butcher recommends, "The first thing we do, let's kill all the lawyers." Is Shakespeare's Dick the Butcher right?

Most legal work is actually not about ambulance chasing, criminal law, and frivolous lawsuits. Instead, it involves the writing and enforcement of contracts, which is how property rights are established. Property rights are essential to protect investments. A good system of laws, by itself, does not provide incentives to invest, because property rights without enforcement are meaningless. This is where lawyers come in. When some-

one encroaches on your land or makes use of your property without your permission, a lawyer can stop him or her. Without lawyers, you would be unwilling to invest. With zero or limited investment, there would be little economic growth.

Canada and the United States have more lawyers per capita than many other countries in the world. They are also among the richest countries in the world with a financial system that is superb at getting capital to new productive uses such as the technology sector. Is this just a coincidence? Or could the legal system actually be beneficial to its economy? Recent research suggests the American legal system, which is based on the Anglo-Saxon legal system, is actually a big advantage for the U.S. economy.*

*See Rafael La Porta, Florencio Lopez-de-Silanes, Andrei Shleifer, and Robert W. Vishny, "Legal Determinants of External Finance," *Journal of Finance* 52 (3), pp. 1131–1150; and Rafael La Porta, Florencio Lopez-de-Silanes, Andrei Shleifer, and Robert W. Vishny, "Law and Finance," *Journal of Political Economy* 106 (6), pp. 1113–1155.

We have seen that government regulation can increase the amount of information in financial markets to make them work more efficiently. Many developing and transition countries have an underdeveloped regulatory apparatus that retards the provision of adequate information to the marketplace. For example, these countries often have weak accounting standards, making it very hard to ascertain the quality of a borrower's balance sheet. As a result, asymmetric information problems are more severe, and the financial system is severely hampered in channelling funds to the most productive uses.

The institutional environment of a poor legal system, weak accounting standards, inadequate government regulation, and government intervention through directed credit programs and state ownership of banks all help explain why many countries stay poor while others grow richer.

APPLICATION**Is China a Counter Example to the Importance of Financial Development?**

Although China appears to be on its way to becoming an economic powerhouse, its financial development is still in its early stages. The country's legal system is weak so that financial contracts are difficult to enforce, while accounting standards are lax so that high-quality information about creditors is hard to find. Regulation of the banking system is still in its formative stages, and the banking sector is dominated by large state-owned banks. Yet China has had one of the highest growth rates in the world over the last twenty years. How has China been able to grow so rapidly given its low level of financial development?

As noted above, China is in an early state of development with a per capita income that is still less than US\$5000, one-eighth that in the United States. With an extremely high savings rate averaging around 40% over the last two decades, it has been able to rapidly build up its capital stock and shift a massive pool of underutilized labour from the subsistence agriculture sector into higher-productivity activities that use capital. Even though available savings have not been allocated to their most productive uses, the huge increase in capital, when combined with the gains in productivity from moving labour out of low-productivity subsistence agriculture, has been enough to produce high growth.

As China gets richer, however, this strategy is unlikely to work. The Soviet Union provides a graphic example. In the 1950s and 60s, the Soviet Union had many similarities to China. It had high growth fuelled by a high savings rate, a massive buildup of capital, and shifts of a large pool of underutilized labour from subsistence agriculture to manufacturing. During this high-growth phase, the Soviet Union was unable to develop the institutions to allocate capital efficiently. As a result, once the pool of subsistence labourers was used up, the Soviet Union's growth slowed dramatically and it was unable to keep up with the West. Today no one considers the Soviet Union to have been an economic success story, and its inability to develop the institutions necessary to sustain financial development and growth was an important reason for the demise of this once superpower.

To get to the next stage of development, China will need to allocate its capital more efficiently, and to do this it has to improve its financial system. The Chinese leadership is well aware of this challenge: The government has announced that state-owned banks are being put on the path to privatization. In addition, the government is engaged in legal reform to make financial contracts more enforceable. New bankruptcy law is being developed so that lenders have the ability to take over the assets of firms that default on their loan contracts. Whether the Chinese government will be successful in developing a first-rate financial system, thereby enabling China to join the ranks of developed countries, is a big question mark.

CONFLICTS OF INTEREST

Earlier in the chapter, we saw how financial institutions play an important role in the financial system. Specifically, their expertise in interpreting signals and collecting information from their customers gives them a cost advantage in the production of information. Furthermore, because they are collecting, producing, and

distributing this information, the financial institutions can use the information over and over again in as many ways as they would like, thereby obtaining economies of scale. By providing multiple financial services to their customers, such as providing them with bank loans or selling their bonds for them, they can also obtain **economies of scope**—that is, they can lower the cost of information production for each service by applying one information resource to many different services. A bank, for example, can evaluate how good a credit risk a corporation is when making them a loan, which then helps the bank decide whether it would be easy for it to sell the bonds of this corporation to the public. Additionally, by providing multiple financial services to their customers, financial institutions develop broader and longer-term relationships with firms. These relationships further reduce the cost of producing information, and further increase economies of scope.

Although the presence of economies of scope may substantially benefit financial institutions, it also creates potential costs in terms of **conflicts of interest**. Conflicts of interest are a type of moral hazard problem that arise when a person or institution has multiple objectives (interests) and, as a result, has conflicts between those objectives. Conflicts of interest are especially likely to occur when a financial institution provides multiple services. The potentially competing interests of those services may lead an individual or firm to conceal information or disseminate misleading information. Here we use the analysis of asymmetric information problems to understand why conflicts of interest are important, why they arise, and what can be done about them.

Why Do We Care About Conflicts of Interest?

We care about conflicts of interest because a substantial reduction in the quality of information in financial markets increases asymmetric information problems and prevents financial markets from channelling funds into productive investment opportunities. Consequently, the financial markets and the economy become less efficient.

Why Do Conflicts of Interest Arise?

Three types of financial service activities have led to prominent conflict-of-interest problems in financial markets in recent years:⁶ underwriting and research in investment banks, auditing and consulting in accounting firms, and credit assessment and consulting in credit-rating agencies. Why do combinations of these activities so often produce conflicts of interest?

UNDERWRITING AND RESEARCH IN INVESTMENT BANKING Investment banks perform two tasks: They *research* companies issuing securities, and they *underwrite* these securities by selling them to the public on behalf of the issuing corporations. Investment banks often combine these distinct financial services because information synergies are possible: That is, information produced for one task may also be useful in the other task. A conflict of interest arises between the brokerage and underwriting services because the banks are attempting to simulta-

⁶ Another important type of conflict of interest arises in universal banking, in which banks engage in multiple financial service activities, including commercial banking, investment banking, and insurance. For further analysis of these conflicts of interest, see Andrew Crockett, Trevor Harris, Frederic S. Mishkin, and Eugene N. White, *Conflicts of Interest in the Financial Services Industry: What Should We Do About Them?*, Geneva Reports on the World Economy 4 (International Center for Monetary and Banking Studies and Centre for Economic Policy Research: Geneva and London, 2003).

neously serve two client groups—the security-issuing firms and the security-buying investors. These client groups have different information needs. Issuers benefit from optimistic research, whereas investors desire unbiased research. However, the same information will be produced for both groups to take advantage of economies of scope. When the potential revenues from underwriting greatly exceed the brokerage commissions from selling, the bank will have a strong incentive to alter the information provided to investors to favour the issuing firm's needs or else risk losing the firm's business to competing investment banks. For example, an internal Morgan Stanley memo excerpted in the *Wall Street Journal* on July 14, 1992, stated, "Our objective . . . is to adopt a policy, fully understood by the entire firm, including the Research Department, that we do not make negative or controversial comments about our clients as a matter of sound business practice."

Because of directives like this one, analysts in investment banks might distort their research to please issuers, and indeed this seems to have happened during the stock market tech boom of the 1990s. Such actions undermine the reliability of the information that investors use to make their financial decisions and, as a result, diminish the efficiency of securities markets.

Another common practice that exploits conflicts of interest is **spinning**. Spinning occurs when an investment bank allocates hot, but underpriced, **initial public offerings (IPOs)**—that is, shares of newly issued stock—to executives of other companies in return for their companies' future business with the investment bank. Because hot IPOs typically immediately rise in price after they are first purchased, spinning is a form of kickback meant to persuade executives to use that investment bank. When the executive's company plans to issue its own shares, he or she will be more likely to go to the investment bank that distributed the hot IPO shares, which is not necessarily the investment bank that would get the highest price for the company's securities. This practice may raise the cost of capital for the firm, thereby diminishing the efficiency of the capital market.

AUDITING AND CONFLICTING INTERESTS IN ACCOUNTING FIRMS Traditionally, an auditor checks the books of companies and monitors the quality of the information produced by firms to reduce the inevitable information asymmetry between the firm's managers and its shareholders. In auditing, threats to truthful reporting arise from several potential conflicts of interest. The conflict of interest that has received the most attention in the media occurs when an accounting firm provides its client with both auditing services and nonaudit consulting services such as advice on taxes, accounting, management information systems, and business strategy. Supplying clients with multiple services allows for economies of scale and scope, but creates two potential sources of conflicts of interest. First, auditors may be willing to skew their judgements and opinions to win consulting business from these same clients. Second, auditors may be auditing information systems or tax and financial plans put in place by their nonaudit counterparts within the firm, and therefore may be reluctant to criticize the systems or advice. Both types of conflicts may lead to biased audits, with the result that less reliable information is available in financial markets and investors find it difficult to allocate capital efficiently.

Another conflict of interest arises when an auditor provides an overly favourable audit to solicit or retain audit business. The unfortunate collapse of Arthur Andersen—once one of the five largest accounting firms in the United States—suggests that this may be the most dangerous conflict of interest.

Credit Assessment and Consulting in Credit-Rating Agencies

Investors use credit ratings (e.g., AAA or BAA) that reflect the probability of default to determine the creditworthiness of particular debt securities. As a consequence, debt ratings play a major role in the pricing of debt securities and in the regulatory process. Conflicts of interest can arise when multiple users with divergent interests (at least in the short term) depend on the credit ratings. Investors and regulators are seeking a well-researched, impartial assessment of credit quality; the issuer needs a favourable rating. In the credit-rating industry, the issuers of securities pay a rating firm such as Standard & Poor's or Moody's to have their securities rated. Because the issuers are the parties paying the credit-rating agency, investors and regulators worry that the agency may bias its ratings upward to attract more business from the issuer.

Another kind of conflict of interest may arise when credit-rating agencies also provide ancillary consulting services. Debt issuers often ask rating agencies to advise them on how to structure their debt issues, usually with the goal of securing a favourable rating. In this situation, the credit-rating agencies would be auditing their own work and would experience a conflict of interest similar to the one found in accounting firms that provide both auditing and consulting services. Furthermore, credit-rating agencies may deliver favourable ratings to garner new clients for the ancillary consulting business. The possible decline in the quality of credit assessments issued by rating agencies could increase asymmetric information in financial markets, thereby diminishing their ability to allocate credit. Such conflicts of interest came to the forefront because of the damaged reputations of the credit-rating agencies during the subprime financial crisis starting in 2007 (see the FYI box, Credit-Rating Agencies and the Subprime Financial Crisis.)

What Has Been Done to Remed Conflicts of Interest?

Two major policy measures were implemented in the United States to deal with conflicts of interest: the Sarbanes-Oxley Act and the Global Legal Settlement.

SARBANES-OXLEY ACT OF 2002 The public outcry over the corporate and accounting scandals in the United States led in 2002 to the passage of the Public Accounting Reform and Investor Protection Act, more commonly referred to as the Sarbanes-Oxley Act, after its two principal authors in Congress. This act increased supervisory oversight to monitor and prevent conflicts of interest:

- It established a Public Company Accounting Oversight Board (PCAOB), overseen by the SEC, to supervise accounting firms and ensure that audits are independent and controlled for quality.
- It increased the SEC's budget to supervise securities markets.

Sarbanes-Oxley also directly reduced conflicts of interest:

- It made it illegal for a registered public accounting firm to provide any non-audit service to a client contemporaneously with an impermissible audit (as determined by the PCAOB).

Sarbanes-Oxley provided incentives for investment banks not to exploit conflicts of interest:

- It beefed up criminal charges for white-collar crime and obstruction of official investigations.

FYI

Credit-Rating Agencies and the Subprime Financial Crisis

Credit-rating agencies have come under severe criticism for the role they played during the subprime financial crisis in the U.S. Credit-rating agencies advised clients on how to structure complex financial instruments that paid out cash flows from subprime mortgages. At the same time, they were rating these identical products, leading to the potential for severe conflicts of interest. Specifically, the large fees they earned from advising clients on how to structure products that they were rating meant they did not have sufficient incentives to make sure their ratings were accurate.

When housing prices began to fall and subprime mortgages began to default, it became crystal clear that the rating agencies had done a terrible job of assessing the risk in the subprime products they had helped to structure. Many AAA-rated products had to be downgraded over and over again until they reached junk status. The resulting massive losses on these assets were one reason why so many financial institutions that were holding them got into trouble, with absolutely disastrous consequences for the economy.

Criticisms of the credit-rating agencies led the U.S. Securities and Exchange Commission (SEC) to propose comprehensive reforms in 2008. The SEC concluded that the credit-

rating agencies' models for rating subprime products were not fully developed and that conflicts of interest may have played a role in producing inaccurate ratings. To address conflicts of interest, the SEC prohibited credit-rating agencies from structuring the same products they rate, prohibited anyone who participates in determining a credit rating from negotiating the fee that the issuer pays for it, and prohibited gifts from bond-issuers to those who rate them in any amount over \$25. In order to make credit-rating agencies more accountable, the SEC's new rules also required more disclosure of how the credit-rating agencies determine ratings. For example, credit-rating agencies were required to disclose historical ratings performance, including the dates of downgrades and upgrades, information on the underlying assets of a product that were used by the credit-rating agencies to rate a product, and the kind of research they used to determine the rating. In addition, the SEC required the rating agencies to differentiate the ratings on structured products from those issued on bonds. The expectation is that these reforms will bring increased transparency to the ratings process and reduce the conflicts of interest that played such a large role in the subprime debacle.

Sarbanes-Oxley also had measures to improve the quality of information in the financial markets:

- It required a corporation's chief executive officer (CEO) and chief financial officer (CFO), as well as its auditors, to certify the accuracy of periodic financial statements and disclosures of the firm (especially regarding off-balance-sheet transactions) (Section 404).
- It required members of the audit committee (the subcommittee of the board of directors that oversees the company's audit) to be "independent"; that is, they cannot be managers in the company or receive any consulting or advisory fee from the company.

FYI

The Demise of Arthur Andersen

In 1913, Arthur Andersen, a young accountant who had denounced the slipshod and deceptive practices that enabled companies to fool the investing public, founded his own firm. Up until the early 1980s, auditing was the most important source of profits within this firm. However, by the late 1980s, the consulting part of the business experienced high revenue growth with high profit margins, while audit profits slumped in a more competitive market. Consulting partners began to assert more power within the firm, and the resulting internal conflicts split the firm in two. Arthur Andersen (the auditing service) and Andersen Consulting were established as separate companies in 2000.

During the period of increasing conflict before the split, Andersen's audit partners had been under increasing pressure to focus on boosting revenue and profits from audit services. Many of Arthur Andersen's clients that later went bust—Enron, WorldCom, Qwest, and Global Crossing—were also the largest clients in Arthur Andersen's regional

offices. The combination of intense pressure to generate revenue and profits from auditing and the fact that some clients dominated regional offices translated into tremendous incentives for regional office managers to provide favourable audit stances for these large clients. The loss of a client like Enron or WorldCom would have been devastating for a regional office and its partners, even if that client contributed only a small fraction of the overall revenue and profits of Arthur Andersen.

The Houston office of Arthur Andersen, for example, ignored problems in Enron's reporting. Arthur Andersen was indicted in March 2002 and then convicted in June 2002 for obstruction of justice for impeding the SEC's investigation of the Enron collapse. Its conviction—the first ever against a major accounting firm—barred Arthur Andersen from conducting audits of publicly traded firms. This development contributed to the firm's demise.

GLOBAL LEGAL SETTLEMENT OF 2002 The second major policy measure arose out of a lawsuit brought by New York Attorney General Eliot Spitzer against the ten largest investment banks (Bear Stearns, Credit Suisse First Boston, Deutsche Bank, Goldman Sachs, J.P. Morgan, Lehman Brothers, Merrill Lynch, Morgan Stanley, Salomon Smith Barney, and UBS Warburg). A global settlement was reached on December 20, 2002, with these investment banks by the SEC, the New York Attorney General, NASD, NASAA, NYSE, and state regulators. Like Sarbanes-Oxley, this settlement directly reduced conflicts of interest:

- It required investment banks to sever the links between research and securities underwriting.
- It banned spinning.

The Global Legal Settlement also provided incentives for investment banks not to exploit conflicts of interest:

- It imposed US\$1.4 billion in fines on the accused investment banks.

The global settlement had measures to improve the quality of information in financial markets:

- It required investment banks to make their analysts' recommendations public.
- Over a five-year period, investment banks were required to contract with at least three independent research firms that would provide research to their brokerage customers.

CONTROL ATTESTATION IN CANADA A great deal of regulatory initiatives with respect to corporate governance have also occupied public attention in Canada in recent years, in reaction to the issues raised by the corporate and accounting scandals in the United States. For example, in October 2002, the Ontario government introduced Bill 198, in response to the strong reforms taking place in the United States. Similar to the Sarbanes-Oxley Act, Bill 198 made several reforms to the securities laws in Ontario, including auditor independence, CEO and CFO accountability for financial reporting, enhanced penalties for illegal activities, and faster disclosure to the public. Moreover, in February 2005 the Canadian Securities Administrators released for comment the Internal Control Instrument and the Certification Instrument, two proposed instruments that substantially mirror the requirements of the Sarbanes-Oxley Act in the United States.

Summar

It is too early to evaluate the impact of the Sarbanes-Oxley Act and the Global Legal Settlement, but the most controversial elements were the separation of functions (research from underwriting, and auditing from nonaudit consulting). Although such a separation of functions may reduce conflicts of interest, it might also diminish economies of scope and thus potentially lead to a reduction of information in financial markets. In addition, there is a serious concern that implementation of these measures, particularly Sarbanes-Oxley, is too costly and is leading to a decline in U.S. capital markets (see the FYI box, *Has Sarbanes-Oxley Led to a Decline in U.S. Capital Markets?*).

SUMMARY

1. There are eight basic facts about financial structure throughout the world. The first four emphasize the importance of financial intermediaries and the relative unimportance of securities markets for the financing of corporations; the fifth recognizes that financial markets are among the most heavily regulated sectors of the economy; the sixth states that only large, well-established corporations have access to securities markets; the seventh indicates that collateral is an important feature of debt contracts; and the eighth presents debt contracts as complicated legal documents that place substantial restrictions on the behaviour of borrowers.
2. Transaction costs freeze many small savers and borrowers out of direct involvement with financial markets. Financial intermediaries can take advantage of economies of scale and are better able to develop expertise to lower transaction costs, thus enabling savers and borrowers to benefit from the existence of financial markets.
3. Asymmetric information results in two problems: adverse selection, which occurs before the transaction, and moral hazard, which occurs after the transaction. Adverse selection refers to the fact that bad credit risks are the ones most likely to seek loans, and moral hazard refers to the risk of the borrower's engaging in activities that are undesirable from the lender's point of view.
4. Adverse selection interferes with the efficient functioning of financial markets. Tools to help reduce the adverse selection problem include private production and sale of information, government regulation to increase information, financial intermediation, and collateral and net worth. The free-rider problem

FYI

Has Sarbanes-Oxley Led to a Decline in U.S. Capital Markets?

There has been much debate in the United States in recent years regarding the impact of Sarbanes-Oxley, especially Section 404, on U.S. capital markets. Section 404 requires both management and company auditors to certify the accuracy of their financial statements. There is no question that Sarbanes-Oxley has led to increased costs for corporations, and this is especially true for smaller firms with revenues of less than US\$100 million, where the compliance costs have been estimated to exceed 1% of sales. These higher costs could result in smaller firms listing abroad and discourage IPOs in the United States, thereby shrinking U.S. capital markets relative to those abroad. However, improved accounting standards could work to encourage stock market listings and IPOs because better information could raise the valuation of common stocks.

Critics of Sarbanes-Oxley have cited it, as well as higher litigation and weaker share-

holder rights, as the cause of declining U.S. stock listings and IPOs, but other factors are likely at work. The European financial system experienced a major liberalization in the 1990s, along with the introduction of the euro, that helped make its financial markets more integrated and efficient. As a result, it became easier for European firms to list in their home countries. The fraction of European firms that list in their home countries has risen to over 90% currently from around 60% in 1995. As the importance of the United States in the world economy has diminished because of the growing importance of other economies, the U.S. capital markets have become less dominant over time. This process is even more evident in the corporate bond market. In 1995, corporate bond issues in the U.S. were double Europe's, while issues of corporate bonds in Europe now exceed those in the United States.

occurs when people who do not pay for information take advantage of information that other people have paid for. This problem explains why financial intermediaries, particularly banks, play a more important role in financing the activities of businesses than securities markets do.

5. Moral hazard in equity contracts is known as the principal-agent problem because managers (the agents) have less incentive to maximize profits than stockholders (the principals). The principal-agent problem explains why debt contracts are so much more prevalent in financial markets than equity contracts. Tools to help reduce the principal-agent problem include monitoring, government regulation to increase information, and financial intermediation.
6. Tools to reduce the moral hazard problem in debt contracts include net worth, monitoring and enforcement of restrictive covenants, and financial intermediaries.
7. Conflicts of interest arise when financial service providers or their employees are serving multiple interests and develop incentives to misuse or conceal information needed for the effective functioning of financial markets. We care about conflicts of interest because they can substantially reduce the amount of reliable information in financial markets, thereby preventing them from channelling funds to those with productive investment opportunities. Two types of financial service activities that have had the greatest potential for conflicts of interest are underwriting and research in investment banking, and auditing and consulting in accounting firms. In the United States, two major policy measures have been implemented to deal with conflicts of interest: the Sarbanes-Oxley Act and the Global Legal Settlement of 2002 arising from the lawsuit by the New York Attorney General against the ten largest investment banks.

KEY TERMS

agency theory, p. 171	free-rider problem, p. 173	restrictive covenants, p. 169
audits, p. 174	incentive-compatible, p. 181	secured debt, p. 169
collateral, p. 169	initial public offerings (IPO), p. 188	spinning, p. 188
conflict of interest, p. 187	net worth (equity capital), p. 176	state-owned banks, p. 185
costly state verification, p. 178	principal-agent problem, p. 177	unsecured debt, p. 169
economies of scope, p. 187		venture capital firm, p. 179

QUESTIONS

You will find the answers to the questions marked with an asterisk in the Textbook Resources section of your MyEconLab.

- How can economies of scale help explain the existence of financial intermediaries?
- *2. Describe two ways in which financial intermediaries help lower transaction costs in the economy.
- Would moral hazard and adverse selection still arise in financial markets if information were not asymmetric? Explain.
- *4. How do standard accounting principles help financial markets work more efficiently?
- Do you think the lemons problem would be more severe for stocks traded on the Toronto Stock Exchange or those traded over the counter? Explain.
- *6. Which firms are most likely to use bank financing rather than to issue bonds or stocks to finance their activities? Why?
- How can the existence of asymmetric information provide a rationale for government regulation of financial markets?
- *8. Would you be more willing to lend to a friend if she put all of her life savings into her business than you would if she had not done so? Why?
- Rich people often worry that others will seek to marry them only for their money. Is this a problem of adverse selection?
- *10. The more collateral there is backing a loan, the less the lender has to worry about adverse selection. Is this statement true, false, or uncertain? Explain your answer.
- How does the free-rider problem aggravate adverse selection and moral hazard problems in financial markets?
- *12. Explain how the separation of ownership and control in Canadian corporations might lead to poor management.
- Why can the provision of several types of financial services by one firm lead to a lower cost of information production?
- *14. How does the provision of several types of financial services by one firm lead to conflicts of interest?
- How can conflicts of interest make financial service firms less efficient?
- *16. Describe two conflicts of interest that occur when underwriting and research are provided by a single investment firm.
- How does spinning lead to a less efficient financial system?
- *18. Describe two conflicts of interest that can occur in accounting firms.
- Which provisions of Sarbanes-Oxley do you think are beneficial, and which do you think are not?
- *20. Which provisions of the Global Legal Settlement do you think are beneficial, and which do you think are not?

WEB EXERCISES

1. In this chapter we discuss the lemons problem and its effect on the efficient functioning of a market. This theory was initially developed by George Akerlof. Go to **www.nobel.se/economics/laureates/2001/public.html**. This site reports that Akerlof, Spence,

and Stiglitz were awarded the Nobel Prize in economics in 2001 for their work. Read this report down through the section on George Akerlof. Summarize his research ideas in one page.



myeconlab

Be sure to visit the MyEconLab website at **www.myeconlab.com**. This online homework and tutorial system puts you in control of your own learning with study and practice tools directly correlated to this chapter content.

CHAPTER 10

Understanding Interest Rates

LEARNING OBJECTIVES

After studying this chapter you should be able to

1. detail the present value concept and the meaning of the term *interest rate*
2. discern among the ways of measuring the interest rate: the yield to maturity, the current yield, and the yield on a discount basis
3. illustrate how bond prices and interest rates are negatively related: when interest rates rise, bond prices fall, and vice versa
4. explain the difference between nominal and real interest rates
5. assess the difference between interest rates and rates of return

PREVIEW

Interest rates are among the most closely watched variables in the economy. Their movements are reported almost daily by the news media because they directly affect our everyday lives and have important consequences for the health of the economy. They affect personal decisions such as whether to consume or save, whether to buy a house, and whether to purchase bonds or put funds into a savings account. Interest rates also affect the economic decisions of businesses and households, such as whether to use their funds to invest in new equipment for factories or to save their money in a bank.

Before we can go on with the study of money, banking, and financial markets, we must understand exactly what the phrase *interest rates* means. In this chapter we see that a concept known as the *yield to maturity* is the most accurate measure of interest rates; the yield to maturity is what economists mean when they use the term *interest rate*. We discuss how the yield to maturity is measured. We'll also see that a bond's interest rate does not necessarily indicate how good an investment the bond is because what it earns (its rate of return) does not necessarily equal its interest rate. Finally, we explore the distinction between real interest rates, which are adjusted for inflation, and nominal interest rates, which are not.

Although learning definitions is not always the most exciting of pursuits, it is important to read carefully and understand the concepts presented in this chapter. Not only are they continually used throughout the remainder of this text, but a firm grasp of these terms will give you a clearer understanding of the role that interest rates play in your life as well as in the general economy.

MEASURING INTEREST RATES

Different debt instruments have very different streams of cash payments to the holder (known as **cash flows**) with very different timing. Thus we first need to understand how we can compare the value of one kind of debt instrument with another before we see how interest rates are measured. To do this, we make use of the concept of *present value*.

Present Value

The concept of **present value** (or **present discounted value**) is based on the commonsense notion that a dollar paid to you one year from now is less valuable to you than a dollar paid to you today: this notion is true because you can deposit a dollar in a savings account that earns interest and have more than a dollar in one year. Economists use a more formal definition, as explained in this section.

Let's look at the simplest kind of debt instrument, which we will call a **simple loan**. In this loan, the lender provides the borrower with an amount of funds (called the *principal*) that must be repaid to the lender at the *maturity date*, along with an additional payment for the interest. For example, if you made your friend, Jane, a simple loan of \$100 for one year, you would require her to repay the principal of \$100 in one year's time along with an additional payment for interest—say, \$10. In the case of a simple loan like this one, the interest payment divided by the amount of the loan is a natural and sensible way to measure the interest rate. This measure of the so-called *simple interest rate*, i , is:

$$i = \frac{\$10}{\$100} = 0.10 = 10\%$$

If you made this \$100 loan, at the end of the year you would have \$110, which can be rewritten as:

$$\$100 \times (1 + 0.10) = \$110$$

If you then lent out the \$110, at the end of the second year you would have:

$$\$110 \times (1 + 0.10) = \$121$$

or, equivalently,

$$\$100 \times (1 + 0.10) \times (1 + 0.10) = \$100 \times (1 + 0.10)^2 = \$121$$

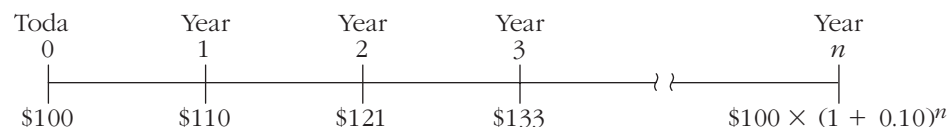
Continuing with the loan again, you would have at the end of the third year:

$$\$121 \times (1 + 0.10) = \$100 \times (1 + 0.10)^3 = \$133$$

Generalizing, we can see that at the end of n years, your \$100 would turn into:

$$\$100 \times (1 + i)^n$$

The amounts you would have at the end of each year by making the \$100 loan today can be seen in the following timeline:



This timeline immediately tells you that you are just as happy having \$100 today as having \$110 a year from now (of course, as long as you are sure that Jane will pay you back). Or that you are just as happy having \$100 today as having \$121 two years from now, or \$133 three years from now, or $\$100 \times (1 + 0.10)^n$ n years from now. The timeline tells us that we can also work backward from future amounts to the present: for example, $\$133 = \$100 \times (1 + 0.10)^3$ three years from now is worth \$100 today, so that:

$$\$100 = \frac{\$133}{(1 + 0.10)^3}$$

APPLICATION

How to Use Your Financial Calculator

The same answer can be obtained by using a financial calculator. Assuming that you have the Texas Instruments BA-35 Solar calculator, set it in the “FIN” mode by pressing the “MODE” key until the word “FIN” appears on the screen, and clear it by pushing the “2nd” key and then the “CE/C” key.

1. Enter 133 and push the “FV” key
2. Enter 10 and push the “%i” key
3. Enter 3 and push the “N” key
4. Enter 0 and push the “PMT” key
5. You want to solve for the present value, so push the “CPT” key and then the “PV” key

The answer is 99.9249.

The process of calculating today’s value of dollars received in the future, as we have done above, is called *discounting the future*. We can generalize this process by writing today’s (present) value of \$100 as PV , the future cash flow (payment) of \$133 as CF , and replacing 0.10 (the 10% interest rate) with i . This leads to the following formula:

$$PV = \frac{CF}{(1 + i)^n} \quad (1)$$

Intuitively, what Equation 1 tells us is that if you are promised \$1 of cash flow for certain ten years from now, this dollar would not be as valuable to you as \$1 is today because if you had the \$1 today, you could invest it and end up with more than \$1 in ten years.

The concept of present value is extremely useful, because it allows us to figure out today’s value (price) of a credit market instrument at a given simple interest rate, i , by just adding up the individual present values of all the future payments received. This information allows us to compare the value of two instruments with very different timing of their payments.

APPLICATION

Simple Present Value

What is the present value of \$250 to be paid in two years if the interest rate is 15%?

Solution

The present value would be \$189.04. Using Equation 1:

$$PV = \frac{CF}{(1 + i)^n}$$

where

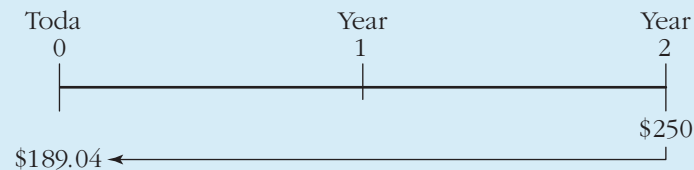
CF = cash flow in two years = \$250

i = annual interest rate = 0.15

n = number of years = 2

Thus

$$PV = \frac{\$250}{(1 + 0.15)^2} = \frac{\$250}{1.3225} = \$189.04$$



To solve using a financial calculator (such as the Texas Instruments BA-35 Solar calculator):

1. Enter 250 and push the “FV” key
2. Enter 15 and push the “%i” key
3. Enter 2 and push the “N” key
4. Enter 0 and push the “PMT” key
5. Push the “CPT” key and then the “PV” key

The answer is 189.0359.

APPLICATION

How Much Is That Jackpot Worth?

As an example of how the present value concept can be used, let's assume that you just hit the \$20 million jackpot in a lottery, which promises you a payment of \$1 million for the next twenty years. You are clearly excited, but have you really won \$20 million?

Solution

No, not in the present value sense. In today's dollars, that \$20 million is worth a lot less. If we assume an interest rate of 10% as in the earlier examples, the first payment of \$1 million is clearly worth \$1 million today, but the next payment next year is only

worth $\$1 \text{ million} / (1 + 0.10) = \$909,091$, a lot less than \$1 million. The following year the payment is worth $\$1 \text{ million} / (1 + 0.10)^2 = \$826,446$ in today's dollars, and so on. When you add all these up, they come to \$9.4 million. You are still pretty excited (who wouldn't be?), but because you understand the concept of present value, you recognize that you are the victim of false advertising. You didn't really win \$20 million, but instead won less than half as much.

Four Types of Credit Market Instruments

In terms of the timing of their cash flow payments, there are four basic types of credit market instruments:

1. A simple loan, which we have already discussed, in which the lender provides the borrower with an amount of funds that must be repaid to the lender at the maturity date along with an additional payment for the interest. Many money market instruments are of this type: for example, commercial loans to businesses.
2. A **fixed-payment loan** (which is also called a **fully amortized loan**) in which the lender provides the borrower with an amount of funds, which must be repaid by making the same payment every period (such as a month) consisting of part of the principal and interest for a set number of years. For example, if you borrowed \$1000, a fixed-payment loan might require you to pay \$126 every year for 25 years. Instalment loans (such as auto loans) and mortgages are frequently of the fixed-payment type.
3. A **coupon bond** pays the owner of the bond a fixed interest payment (coupon payment) every year until the maturity date, when a specified final amount (**face value** or **par value**) is repaid. The coupon payment is so named because the bondholder used to obtain payment by clipping a coupon off the bond and sending it to the bond issuer, who then sent the payment to the holder. Nowadays, it is no longer necessary to send in coupons to receive these payments. A coupon bond with \$1000 face value, for example, might pay you a coupon payment of \$100 per year for ten years and at the maturity date repay you the face value amount of \$1000. (The face value of a bond is usually in \$1000 increments.)

A coupon bond is identified by three pieces of information. First is the corporation or government agency that issues the bond. Second is the maturity date of the bond. Third is the bond's **coupon rate**, the dollar amount of the yearly coupon payment expressed as a percentage of the face value of the bond. In our example, the coupon bond has a yearly coupon payment of \$100 and a face value of \$1000. The coupon rate is then $\$100 / \$1000 = 0.10$, or 10%. Canada bonds and corporate bonds are examples of coupon bonds.

4. A **discount bond** (also called a **zero-coupon bond**) is bought at a price below its face value (at a discount), and the face value is repaid at the maturity date. Unlike a coupon bond, a discount bond does not make any interest payments; it just pays off the face value. For example, a discount bond with a face value of \$1000 might be bought for \$900 and in a year's time the owner would be repaid the face value of \$1000. Canadian government treasury bills and long-term zero-coupon bonds are examples of discount bonds.

These four types of instruments require payments at different times: simple loans and discount bonds make payment only at their maturity dates, whereas fixed-payment loans and coupon bonds have payments periodically until maturity.

How would you decide which of these instruments provides you with more income? They all seem so different because they make payments at different times. To solve this problem, we use the concept of present value to provide us with a procedure for measuring interest rates on these different types of instruments.

Yield to Maturity

Of the several common ways of calculating interest rates, the most important is the **yield to maturity**, the interest rate that equates the present value of cash flow payments received from a debt instrument with its value today.¹ Because the concept behind the calculation of yield to maturity makes good economic sense, economists consider it the most accurate measure of interest rates.

To understand yield to maturity better, we now look at how it is calculated for the four types of credit market instruments. In all these examples, the key to understanding the calculation of the yield to maturity is equating today's value of the debt instrument with the present value of all of its future cash flows.

SIMPLE LOAN Using the concept of present value, the yield to maturity on a simple loan is easy to calculate. For the one-year loan we discussed, today's value is \$100, and the payments in one year's time would be \$110 (the repayment of \$100 plus the interest payment of \$10). We can use this information to solve for the yield to maturity i by recognizing that the present value of the future payments must equal today's value of a loan.

APPLICATION

Yield to Maturity on a Simple Loan

If Pete borrows \$100 from his sister and next year she wants \$110 back from him, what is the yield to maturity on this loan?

Solution

The yield to maturity on the loan is 10%.

$$PV = \frac{CF}{(1 + i)^n}$$

where

$$PV = \text{amount borrowed} = \$100$$

$$CF = \text{cash flow in one year} = \$110$$

$$n = \text{number of years} = 1$$

Thus

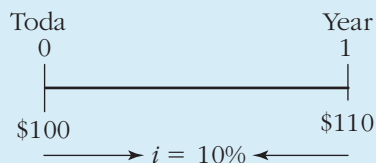
$$\$100 = \frac{\$110}{(1 + i)}$$

$$(1 + i) \$100 = \$110$$

$$(1 + i) = \frac{\$110}{\$100}$$

$$i = 1.10 - 1 = 0.10 = 10\%$$

¹ In other contexts, it is also called the *internal rate of return*.



To find the yield to maturity using a financial calculator:

1. Enter 100 and push the “PV” key
2. Enter 110 and push the “FV” key
3. Enter 1 and push the “N” key
4. Enter 0 and push the “PMT” key
5. Push the “CPT” key and then the “%i” key

The answer is 10.

This calculation of the yield to maturity should look familiar because it equals the interest payment of \$10 divided by the loan amount of \$100; that is, it equals the simple interest rate on the loan. An important point to recognize is that **for simple loans, the simple interest rate equals the yield to maturity**. Hence the same term i is used to denote both the yield to maturity and the simple interest rate.

FIXED-PAYMENT LOAN Recall that this type of loan has the same cash flow payment every period throughout the life of the loan. On a fixed-rate mortgage, for example, the borrower makes the same payment to the bank every month until the maturity date, when the loan will be completely paid off. To calculate the yield to maturity for a fixed-payment loan, we follow the same strategy we used for the simple loan—we equate today’s value of the loan with its present value. Because the fixed-payment loan involves more than one cash flow payment, the present value of the fixed-payment loan is calculated as the sum of the present values of all payments (using Equation 1).

In the case of our earlier example, the loan is \$1000 and the yearly cash flow payment is \$126 for the next 25 years. The present value is calculated as follows: at the end of one year there is a \$126 payment with a PV of $\$126/(1 + i)$; at the end of two years there is another \$126 payment with a PV of $\$126/(1 + i)^2$; and so on until at the end of the twenty-fifth year, the last payment of \$126 with a PV of $\$126/(1 + i)^{25}$ is made. Making today’s value of the loan (\$1000) equal to the sum of the present values of all the yearly payments gives us

$$\$1000 = \frac{\$126}{1 + i} + \frac{\$126}{(1 + i)^2} + \frac{\$126}{(1 + i)^3} + \cdots + \frac{\$126}{(1 + i)^{25}}$$

More generally, for any fixed-payment loan,

$$PV = \frac{FP}{1 + i} + \frac{FP}{(1 + i)^2} + \frac{FP}{(1 + i)^3} + \cdots + \frac{FP}{(1 + i)^n} \quad (2)$$

where

IV = loan value

FP = fixed yearly payment

n = number of years until maturity

For a fixed-payment loan amount, the fixed yearly payment and the number of years until maturity are known quantities, and only the yield to maturity is not. So we can solve this equation for the yield to maturity i . Because this calculation is not easy, many pocket calculators have programs that allow you to find i given the loan's numbers for IV , FP , and n . For example, in the case of the 25-year loan with yearly payments of \$126, the yield to maturity that solves Equation 2 is 12%. Real estate brokers always have a pocket calculator that can solve such equations so that they can immediately tell the prospective house buyer exactly what the yearly (or monthly) payments will be if the house purchase is financed by taking out a mortgage.

APPLICATION

Yield to Maturity on a Fixed-Payment Loan

You decide to purchase a new home and need a \$100 000 mortgage. You take out a loan from the bank that has an interest rate of 7%. What is the yearly payment to the bank to pay off the loan in 20 years?

Solution

The yearly payment to the bank is \$9439.29.

$$IV = \frac{FP}{1+i} + \frac{FP}{(1+i)^2} + \frac{FP}{(1+i)^3} + \cdots + \frac{FP}{(1+i)^n}$$

where

IV = loan value amount = \$100 000

i = annual interest rate = 0.07

n = number of years = 20

Thus

$$\$100\,000 = \frac{FP}{1+0.07} + \frac{FP}{(1+0.07)^2} + \frac{FP}{(1+0.07)^3} + \cdots + \frac{FP}{(1+0.07)^{20}}$$

To find the yearly payment for the loan using a financial calculator:

1. Enter -100 000 and push the "PV" key
2. Enter 0 and push the "FV" key
3. Enter 20 and push the "N" key
4. Enter 7 and push the "%i" key
5. Push the "CPT" and "PMT" keys

The answer is 9439.29.

COUPON BOND To calculate the yield to maturity for a coupon bond, follow the same strategy used for the fixed-payment loan: equate today's value of the bond with its present value. Because coupon bonds also have more than one cash flow payment, the present value of the bond is calculated as the sum of the present values of all the coupon payments plus the present value of the final payment of the face value of the bond.

The present value of a \$1000-face-value bond with ten years to maturity and yearly coupon payments of \$100 (a 10% coupon rate) can be calculated as follows: at the end of one year, there is a \$100 coupon payment with a *PV* of $\$100/(1+i)$; at the end of the second year, there is another \$100 coupon payment with a *PV* of $\$100/(1+i)^2$; and so on until, at maturity, there is a \$100 coupon payment with a *PV* of $\$100/(1+i)^{10}$ plus the repayment of the \$1000 face value with a *PV* of $\$1000/(1+i)^{10}$. Setting today's value of the bond (its current price, denoted by *P*) equal to the sum of the present values of all the cash flow payments for this bond gives

$$P = \frac{\$100}{1+i} + \frac{\$100}{(1+i)^2} + \frac{\$100}{(1+i)^3} + \cdots + \frac{\$100}{(1+i)^{10}} + \frac{\$1000}{(1+i)^{10}}$$

More generally, for any coupon bond,²

$$P = \frac{C}{1+i} + \frac{C}{(1+i)^2} + \frac{C}{(1+i)^3} + \cdots + \frac{C}{(1+i)^n} + \frac{F}{(1+i)^n} \quad (3)$$

where

P = price of coupon bond
C = yearly coupon payment
F = face value of the bond
n = years to maturity date

In Equation 3, the coupon payment, the face value, the years to maturity, and the price of the bond are known quantities, and only the yield to maturity is not. Hence we can solve this equation for the yield to maturity *i*. Just as in the case of the fixed-payment loan, this calculation is not easy, so business-oriented pocket calculators have built-in programs that solve this equation for you.

APPLICATION

Yield to Maturity on a Coupon Bond

Find the price of a 10% coupon bond with a face value of \$1000, a 12.25% yield to maturity, and eight years to maturity.

Solution

To solve using the Texas Instruments BA-35 Solar calculator:

1. Enter 1000 and push the “FV” key
2. Enter 8 and push the “N” key
3. Enter 12.25 and push the “%i” key
4. Enter 100 and push the “PMT” key
5. Push the “CPT” and “PV” keys

The answer is 889.1977.

² Most coupon bonds actually make coupon payments on a semi-annual basis rather than once a year as assumed here. The effect on the calculations is only very slight and will be ignored here.

Table 4-1 shows the yields to maturity calculated for several bond prices. Three interesting facts emerge:

1. When the coupon bond is priced at its face value, the yield to maturity equals the coupon rate.
2. The price of a coupon bond and the yield to maturity are negatively related; that is, as the yield to maturity rises, the price of the bond falls. As the yield to maturity falls, the price of the bond rises.
3. The yield to maturity is greater than the coupon rate when the bond price is below its face value.

These three facts are true for any coupon bond and are really not surprising if you think about the reasoning behind the calculation of the yield to maturity. When you put \$1000 in a bank account with an interest rate of 10%, you can take out \$100 every year and you will be left with the \$1000 at the end of ten years. This is similar to buying the \$1000 bond with a 10% coupon rate analyzed in Table 4-1, which pays a \$100 coupon payment every year and then repays \$1000 at the end of ten years. If the bond is purchased at the par value of \$1000, its yield to maturity must equal 10%, which is also equal to the coupon rate of 10%. The same reasoning applied to any coupon bond demonstrates that if the coupon bond is purchased at its par value, the yield to maturity and the coupon rate must be equal.

It is straightforward to show that the bond price and the yield to maturity are negatively related. As i , the yield to maturity, rises, all denominators in the bond price formula must necessarily rise. Hence a rise in the interest rate as measured by the yield to maturity means that the price of the bond must fall. Another way to explain why the bond price falls when the interest rate rises is that a higher interest rate implies that the future coupon payments and final payment are worth less when discounted back to the present; hence the price of the bond must be lower.

The third fact, that the yield to maturity is greater than the coupon rate when the bond price is below its par value, follows directly from facts 1 and 2. When the yield to maturity equals the coupon rate, then the bond price is at the face value, and when the yield to maturity rises above the coupon rate, the bond price necessarily falls and so must be below the face value of the bond.

There is one special case of a coupon bond that is worth discussing because its yield to maturity is particularly easy to calculate. This bond is called a **consol** or a **perpetuit**; it is a perpetual bond with no maturity date and no repayment of principal that makes fixed coupon payments of $\$C$ forever. Consols were first sold by the British Treasury during the Napoleonic Wars and are still traded today;

TABLE 10-1 Yields to Maturity on a 10% Coupon-Rate Bond Maturing in Ten Years (Face Value = \$1000)

Price of Bond (\$)	Yield to Maturity (%)
1200	7.13
1100	8.48
1000	10.00
900	11.75
800	13.81

they are quite rare, however, in Canadian capital markets. The formula in Equation 3 for the price of the consol P simplifies to the following:³

$$P_c = \frac{C}{i_c} \quad (4)$$

where P_c = price of the perpetuity (consol)
 C = yearly payment
 i_c = yield to maturity of the perpetuity (consol)

One nice feature of perpetuities is that you can immediately see that as i_c goes up, the price of the bond falls. For example, if a perpetuity pays \$100 per year forever and the interest rate is 10%, its price will be \$1000 = \$100/0.10. If the interest rate rises to 20%, its price will fall to \$500 = \$100/0.20. We can also rewrite this formula as

$$i_c = \frac{C}{P_c} \quad (5)$$

The formula in Equation 5, which describes the calculation of the yield to maturity for a perpetuity, also provides a useful approximation for the yield to maturity on coupon bonds. When a coupon bond has a long term to maturity (say, 20 years or more), it is very much like a perpetuity, which pays coupon payments forever. This is because the cash flows more than 20 years in the future have such small present discounted values that the value of a long-term coupon bond is very close to the value of a perpetuity with the same coupon rate. Thus i_c in Equation 5 will be very close to the yield to maturity for any long-term bond. For this reason, i_c , the yearly coupon payment divided by the price of the security, has been given the name **current yield** and is frequently used as an approximation to describe interest rates on long-term bonds.

DISCOUNT BOND The yield-to-maturity calculation for a discount bond is similar to that for the simple loan. Let us consider a discount bond such as a one-year Canadian treasury bill, which pays off a face value of \$1000 in one year's time. If the current purchase price of this bill is \$900, then equating this price to the pre-

³ The bond price formula for a consol is

$$P = \frac{C}{1+i} + \frac{C}{(1+i)^2} + \frac{C}{(1+i)^3} + \dots$$

which can be written as

$$P = C(x + x^2 + x^3 + \dots)$$

in which $x = 1/(1+i)$. The formula for an infinite sum is:

$$1 + x + x^2 + x^3 + \dots = \frac{1}{1-x} \text{ for } x < 1$$

and so

$$P = C \left(\frac{1}{1-x} - 1 \right) = C \left(\frac{1}{1-1/(1+i)} - 1 \right)$$

which by suitable algebraic manipulation becomes

$$P = C \left(\frac{1+i}{i} - \frac{i}{i} \right) = \frac{C}{i}$$

APPLICATION

Yield to Maturity on a Perpetuity

What is the yield to maturity on a bond that has a price of \$2000 and pays \$100 annually forever?

Solution

The yield to maturity would be 5%.

$$i_c = \frac{C}{P_c}$$

where

$$C = \text{yearly payment} = \$100$$

$$P_c = \text{price of perpetuity (consol)} = \$2000$$

Thus

$$i_c = \frac{\$100}{\$2000}$$

$$i_c = 0.05 = 5\%$$

To solve using the Texas Instruments BA-35 Solar calculator:

1. Enter 2000 and push the “PV” key
2. Enter 0 and push the “FV” key
3. Enter 999 (to approximate an infinite number of payments) and push the “N” key
4. Enter 100 and push the “PMT” key
5. Push the “CPT” and “%i” keys

The answer is 5.

present value of the \$1000 received in one year, using Equation 1 (page 60), gives

$$\$900 = \frac{\$1000}{1 + i}$$

and solving for i ,

$$(1 + i) \times \$900 = \$1000$$

$$\$900 + \$900i = \$1000$$

$$\$900i = \$1000 - \$900$$

$$i = \frac{\$1000 - \$900}{\$900} = 0.111 = 11.1\%$$

More generally, for any one-year discount bond, the yield to maturity can be written as

$$i = \frac{F - P}{P} \quad (6)$$

where

F = face value of the discount bond
 P = current price of the discount bond

In other words, the yield to maturity equals the increase in price over the year, $F - P$, divided by the initial price P . In normal circumstances, investors earn positive returns from holding these securities and so they sell at a discount, meaning that the current price of the bond is below the face value. Therefore, $F - P$ should be positive, and the yield to maturity should be positive as well. However, this is not always the case, as recent extraordinary events in Japan indicate (see the Global box, Negative T-Bill Rates? Japan Shows the Way).

An important feature of this equation is that it indicates that for a discount bond, the yield to maturity is negatively related to the current bond price. This is the same conclusion that we reached for a coupon bond. For example, Equation 6 shows that a rise in the bond price from \$900 to \$950 means that the bond will have a smaller increase in its price at maturity, and the yield to maturity falls from 11.1% to 5.3%. Similarly, a fall in the yield to maturity means that the price of the discount bond has risen.

SUMMARY The concept of present value tells you that a dollar in the future is not as valuable to you as a dollar today because you can earn interest on this dollar. Specifically, a dollar received n years from now is worth only $\$1/(1 + i)^n$

GLOBAL

Negative T-Bill Rates? Japan Shows the Way

We normally assume that interest rates must always be positive. Negative interest rates would imply that you are willing to pay more for a bond today than you will receive for it in the future (as our formula for yield to maturity on a discount bond demonstrates). Negative interest rates therefore seem like an impossibility because you would do better by holding cash that has the same value in the future as it does today.

The Japanese have demonstrated that this reasoning is not quite correct. In November 1998, interest rates on Japanese six-month treasury bills became negative, yielding an interest rate of -0.004% , with investors paying more for the bills than their face value. This is an extremely unusual event because no other country in the world has seen negative

interest rates during the last fifty years. How could this happen?

As we will see in Chapter 5, the weakness of the Japanese economy and a negative inflation rate drove Japanese interest rates to low levels, but these two factors can't explain the negative rates. The answer is that large investors found it more convenient to hold these six-month bills as a store of value rather than holding cash because the bills are denominated in larger amounts and can be stored electronically. For that reason, some investors were willing to hold them, despite their negative rates, even though in monetary terms the investors would be better off holding cash. Clearly, the convenience of T-bills only goes so far, and thus their interest rates can go only a little bit below zero.

today. The present value of a set of future cash flow payments on a debt instrument equals the sum of the present values of each of the future payments. The yield to maturity for an instrument is the interest rate that equates the present value of the future payments on that instrument to its value today. Because the procedure for calculating yield to maturity is based on sound economic principles, this is the measure that economists think most accurately describes the interest rate.

Our calculations of the yield to maturity for a variety of bonds reveal the important fact that **current bond prices and interest rates are negatively related: when the interest rate rises, the price of the bond falls, and vice versa.**

THE DISTINCTION BETWEEN INTEREST RATES AND RETURNS

Many people think that the interest rate on a bond tells them all they need to know about how well off they are as a result of owning it. If Irving the Investor thinks he is better off when he owns a long-term bond yielding a 10% interest rate and the interest rate rises to 20%, he will have a rude awakening: as we will see shortly, if he has to sell the bond, Irving has lost his shirt! How well a person does by holding a bond or any other security over a particular time period is accurately measured by the **rate of return** or, in more precise terminology, the **rate of return**. The concept of **rate of return** discussed here is extremely important because it is used continually throughout this book and understanding it will make the material presented later in the book easier to follow. For any security, the rate of return is defined as the payments to the owner plus the change in its value, expressed as a fraction of its purchase price. To make this definition clearer, let us see what the return would look like for a \$1000-face-value coupon bond with a coupon rate of 10% that is bought for \$1000, held for one year, and then sold for \$1200. The payments to the owner are the yearly coupon payments of \$100, and the change in its value is \$1200 - \$1000 = \$200. Adding these together and expressing them as a fraction of the purchase price of \$1000 gives us the one-year holding-period return for this bond:

$$\frac{\$100 + \$200}{\$1000} = \frac{\$300}{\$1000} = 0.30 = 30\%$$

You may have noticed something quite surprising about the return that we have just calculated: it equals 30%, yet as Table 4-1 (page 67) indicates, initially the yield to maturity was only 10%. This demonstrates that **the return on a bond will not necessarily equal the yield to maturity on that bond.** We now see that the distinction between interest rate and return can be important, although for many securities the two may be closely related.

More generally, the return on a bond held from time t to time $t + 1$ can be written as

$$RET = \frac{C + P_{t+1} - P_t}{P_t} \quad (8)$$

where

RET = return from holding the bond from time t to time $t + 1$

P_t = price of the bond at time t

P_{t+1} = price of the bond at time $t + 1$

C = coupon payment

APPLICATION

Calculating the Rate of Return

What would the rate of return be on a bond bought for \$1000 and sold one year later for \$800? The bond has a face value of \$1000 and a coupon rate of 8%.

Solution

The rate of return on the bond for holding it one year is -12% .

$$RET = \frac{C + P_{+1} - P}{P}$$

where

$$C = \text{coupon payment} = \$1000 \times 0.08 = \$80$$

$$P_{+1} = \text{price of the bond one year later} = \$800$$

$$P = \text{price of the bond today} = \$1000$$

Thus

$$RET = \frac{\$80 + (\$800 - \$1000)}{\$1000} = \frac{-\$120}{\$1000} = -0.12 = -12\%$$

A convenient way to rewrite the return formula in Equation 8 is to recognize that it can be split into two separate terms:

$$RET = \frac{C}{P} + \frac{P_{+1} - P}{P}$$

The first term is the current yield i_c (the coupon payment over the purchase price):

$$\frac{C}{P} = i_c$$

The second term is the **rate of capital gain**, or the change in the bond's price relative to the initial purchase price:

$$\frac{P_{+1} - P}{P} = g$$

where g = rate of capital gain. Equation 8 can then be rewritten as

$$RET = i_c + g \quad (9)$$

which shows that the return on a bond is the current yield i_c plus the rate of capital gain g . This rewritten formula illustrates the point we just discovered. Even for a bond for which the current yield i_c is an accurate measure of the yield to maturity, the return can differ substantially from the interest rate. Returns will differ from the interest rate especially if there are sizable fluctuations in the price of the bond that produce substantial capital gains or losses.

APPLICATION

Calculating the Rate of Capital Gain

Calculate the rate of capital gain or loss on a ten-year zero-coupon bond for which the interest rate has increased from 10% to 20%. The bond has a face value of \$1000.

Solution

The rate of capital gain or loss is -49.7% .

$$g = \frac{P_{+1} - P}{P}$$

where

$$P_{+1} = \text{price of the bond one year from now} = \frac{\$1000}{(1 + 0.20)^9} = \$193.81$$

$$P = \text{price of the bond today} = \frac{\$1000}{(1 + 0.10)^{10}} = \$385.54$$

Thus

$$g = \frac{\$193.81 - \$385.54}{\$385.54}$$

$$g = -0.497 = -49.7\%$$

To explore this point even further, let's look at what happens to the returns on bonds of different maturities when interest rates rise. Table 4-2 calculates the one-year return using Equation 9 on several 10%-coupon-rate bonds all purchased at par when interest rates on all these bonds rise from 10% to 20%. Several key findings in this table are generally true of all bonds:

- The only bond whose return equals the initial yield to maturity is one whose time to maturity is the same as the holding period (see the last bond in Table 4-2).
- A rise in interest rates is associated with a fall in bond prices, resulting in capital losses on bonds whose terms to maturity are longer than the holding period.
- The more distant a bond's maturity, the greater the size of the percentage price change associated with an interest-rate change.
- The more distant a bond's maturity, the lower the rate of return that occurs as a result of the increase in the interest rate.
- Even though a bond has a substantial initial interest rate, its return can turn out to be negative if interest rates rise.

At first it frequently puzzles students (as it puzzles poor Irving the Investor) that a rise in interest rates can mean that a bond has been a poor investment. The trick to understanding this is to recognize that a rise in the interest rate means that the price of a bond has fallen. A rise in interest rates therefore means that a capital loss has occurred, and if this loss is large enough, the bond can be a poor

TABLE 10-2 One-Year Returns on Different-Maturity 10%-Coupon-Rate Bonds When Interest Rates Rise from 10% to 20%

(1) Years to Maturity When Bond Is Purchased	(2) Initial Current Yield (%)	(3) Initial Price (\$)	(4) Price Next Year* (\$)	(5) Rate of Capital Gain (%)	(6) Rate of Return (2 + 5) (%)
30	10	1000	503	−49.7	−39.7
20	10	1000	516	−48.4	−38.4
10	10	1000	597	−40.3	−30.3
5	10	1000	741	−25.9	−15.9
2	10	1000	917	−8.3	+1.7
1	10	1000	1000	0.0	+10.0

*Calculated with a financial calculator using Equation 3.

investment indeed. For example, we see in Table 4-2 that the bond that has 30 years to maturity when purchased has a capital loss of 49.7% when the interest rate rises from 10% to 20%. This loss is so large that it exceeds the current yield of 10%, resulting in a negative return (loss) of −39.7%. If Irving does not sell the bond, his capital loss is often referred to as a “paper loss.” This is a loss nonetheless because if he had not bought this bond and had instead put his money in the bank, he would now be able to buy more bonds at their lower price than he presently owns.

Maturity and the Volatility of Bond Returns: Interest-Rate Risk

The finding that the prices of longer-maturity bonds respond more dramatically to changes in interest rates helps explain an important fact about the behaviour of bond markets: **prices and returns for long-term bonds are more volatile than those for shorter-term bonds.** Price changes of +20% and −20% within a year, with corresponding variations in returns, are common for bonds more than 20 years away from maturity.

We now see that changes in interest rates make investments in long-term bonds quite risky. Indeed, the riskiness of an asset's return that results from interest-rate changes is so important that it has been given a special name, **interest-rate risk**.⁴ Dealing with interest-rate risk is a major concern of managers of financial institutions and investors, as we will see in later chapters (see also the FYI box Helping Investors to Select Desired Interest-Rate Risk).

Although long-term debt instruments have substantial interest-rate risk, short-term debt instruments do not. Indeed, bonds with a maturity that is as short as the

⁴ Interest-rate risk can be quantitatively measured using the concept of duration. This concept and how it is calculated are discussed in an appendix to this chapter, which can be found on this book's MyEconLab www.pearsoned.ca/m_econlab.

FYI

Helping Investors to Select Desired Interest-Rate Risk

Because many investors want to know how much interest-rate risk they are exposed to, some mutual fund companies try to educate investors about the perils of interest-rate risk, as well as to offer investment alternatives that match their investors' preferences.

For example, one U.S. company, Vanguard Group, offers eight separate high-grade bond mutual funds. In its prospectus, Vanguard separates the funds by the average maturity of the bonds they hold and demonstrates the effect of interest-rate changes by computing the percentage change in bond value resulting from a 1% increase and decrease in interest rates.

Three of the bond funds invest in bonds with average maturities of one to three years, which Vanguard rates as having the lowest interest-rate risk. Three other funds hold bonds with average maturities of five to ten years, which Vanguard rates as having medium interest-rate risk. Two funds hold long-term bonds with maturities of 15 to 30 years, which Vanguard rates as having high interest-rate risk.

By providing this information, Vanguard hopes to increase its market share in the sale of bond funds. Not surprisingly, Vanguard is one of the most successful mutual fund companies in the business.

holding period have no interest-rate risk.⁵ We see this for the coupon bond at the bottom of Table 4-2, which has no uncertainty about the rate of return because it equals the yield to maturity, which is known at the time the bond is purchased. The key to understanding why there is no interest-rate risk for any bond whose time to maturity matches the holding period is to recognize that (in this case) the price at the end of the holding period is already fixed at the face value. The change in interest rates can then have no effect on the price at the end of the holding period for these bonds, and the return will therefore be equal to the yield to maturity known at the time the bond is purchased.⁶

⁵ The statement that there is no interest-rate risk for any bond whose time to maturity matches the holding period is literally true only for discount bonds and zero-coupon bonds that make no intermediate cash payments before the holding period is over. A coupon bond that makes an intermediate cash payment before the holding period is over requires that this payment be reinvested. Because the interest rate at which this payment can be reinvested is uncertain, there is some uncertainty about the return on this coupon bond even when the time to maturity equals the holding period. However, the riskiness of the return on a coupon bond from reinvesting the coupon payments is typically quite small, and so the basic point that a coupon bond with a time to maturity equalling the holding period has very little risk still holds true.

⁶ In the text, we are assuming that all holding periods are short and equal to the maturity on short-term bonds and are thus not subject to interest-rate risk. However, if an investor's holding period is longer than the term to maturity of the bond, the investor is exposed to a type of interest-rate risk called *reinvestment risk*. Reinvestment risk occurs because the proceeds from the short-term bond need to be reinvested at a future interest rate that is uncertain.

To understand reinvestment risk, suppose that Irving the Investor has a holding period of two years and decides to purchase a \$1000 one-year bond at face value and will then purchase another one at the end of the first year. If the initial interest rate is 10%, Irving will have \$1100 at the end of the year. If the interest rate rises to 20%, as in Table 4-2, Irving will find that buying \$1100 worth of another one-year bond will leave him at the end of the second year with $\$1100 \times (1 + 0.20) = \1320 . Thus Irving's two-year return will be $(\$1320 - \$1000)/\$1000 = 0.32 = 32\%$, which equals 14.9% at an annual rate. In this case, Irving has earned more by buying the one-year bonds than if he had initially purchased the two-year bond with an interest rate of 10%. Thus when Irving has a holding period that is longer than the

Summary

The return on a bond, which tells you how good an investment it has been over the holding period, is equal to the yield to maturity in only one special case: when the holding period and the maturity of the bond are identical. Bonds whose term to maturity is longer than the holding period are subject to interest-rate risk: changes in interest rates lead to capital gains and losses that produce substantial differences between the return and the yield to maturity known at the time the bond is purchased. Interest-rate risk is especially important for long-term bonds, where the capital gains and losses can be substantial. This is why long-term bonds are not considered to be safe assets with a sure return over short holding periods.

THE DISTINCTION BETWEEN REAL AND NOMINAL INTEREST RATES

So far in our discussion of interest rates, we have ignored the effects of inflation on the cost of borrowing. What we have up to now been calling the interest rate makes no allowance for inflation, and it is more precisely referred to as the **nominal interest rate**. We distinguish it from the **real interest rate**, the interest rate that is adjusted by subtracting expected changes in the price level (inflation) so that it more accurately reflects the true cost of borrowing. This interest rate is more precisely referred to as the *ex ante real interest rate* because it is adjusted for *expected* changes in the price level. The *ex ante* real interest rate is most important to economic decisions, and typically it is what economists mean when they make reference to the “real” interest rate. The interest rate that is adjusted for *actual* changes in the price level is called the *ex post real interest rate*. It describes how well a lender has done in real terms after the fact.

The real interest rate is more accurately defined by the *Fisher equation*, named for Irving Fisher, one of the great monetary economists of the twentieth century. The Fisher equation states that the nominal interest rate i equals the real interest rate i_r plus the expected rate of inflation π^e .⁷

$$i = i_r + \pi^e \quad (10)$$

⁶ (*continued*) term to maturity of the bonds he purchases, he benefits from a rise in interest rates. Conversely, if interest rates fall to 5%, Irving will have only \$1155 at the end of two years: $\$1100 \times (1 + 0.05)$. Thus his two-year return will be $(\$1155 - \$1000)/\$1000 = 0.155 = 15.5\%$, which is 7.5% at an annual rate. With a holding period greater than the term to maturity of the bond, Irving now loses from a fall in interest rates.

We have thus seen that when the holding period is longer than the term to maturity of a bond, the return is uncertain because the future interest rate when reinvestment occurs is also uncertain—in short, there is reinvestment risk. We also see that if the holding period is longer than the term to maturity of the bond, the investor benefits from a rise in interest rates and is hurt by a fall in interest rates.

⁷ A more precise formulation of the Fisher equation is

$$i = i_r + \pi^e + (i_r \times \pi^e)$$

because

$$1 + i = (1 + i_r)(1 + \pi^e) = 1 + i_r + \pi^e + (i_r \times \pi^e)$$

and subtracting 1 from both sides gives us the first equation. For small values of i_r and π^e , the term $i_r \times \pi^e$ is so small that we ignore it, as in the text.

Rearranging terms, we find that the real interest rate equals the nominal interest rate minus the expected inflation rate:

$$i_r = i - \pi^e \quad (11)$$

To see why this definition makes sense, let us first consider a situation in which you have made a one-year simple loan with a 5% interest rate ($i = 5\%$) and you expect the price level to rise by 3% over the course of the year ($\pi^e = 3\%$). As a result of making the loan, at the end of the year you expect to have 2% more in **real erms**, that is, in terms of real goods and services you can buy. In this case, the interest rate you expect to earn in terms of real goods and services is 2%; that is,

$$i_r = 5\% - 3\% = 2\%$$

as indicated by the Fisher definition.

A similar distinction can be made between nominal returns and real returns. Nominal returns, which do not allow for inflation, are what we have been referring to as simply “returns.” When inflation is subtracted from a nominal return, we have the real return, which indicates the amount of extra goods and services that can be purchased as a result of holding the security.

The distinction between real and nominal interest rates is important because the real interest rate, which reflects the real cost of borrowing, is likely to be a better indicator of the incentives to borrow and lend. It appears to be a better guide to how people will be affected by what is happening in credit markets. Figure 4-1,

APPLICATION

Calc lating Real Interest Rates

What is the real interest rate if the nominal interest rate is 8% and the expected inflation rate is 10% over the course of a year?

Sol tion

The real interest rate is -2% . Although you will be receiving 8% more dollars at the end of the year, you will be paying 10% more for goods. The result is that you will be able to buy 2% fewer goods at the end of the year, and you will be 2% worse off in real terms.

$$i_r = i - \pi^e$$

where

$$i = \text{nominal interest rate} = 0.08$$

$$\pi^e = \text{expected inflation rate} = 0.10$$

Thus

$$i_r = 0.08 - 0.10 = -0.02 = -2\%$$

As a lender, you are clearly less eager to make a loan in this case because in terms of real goods and services you have actually earned a negative interest rate of 2%. By contrast, as the borrower, you fare quite well because at the end of the year, the amounts you will have to pay back will be worth 2% less in terms of goods and services—you as the borrower will be ahead by 2% in real terms. **When the real interest rate is low, there are greater incentives to borrow and fewer incentives to lend.**

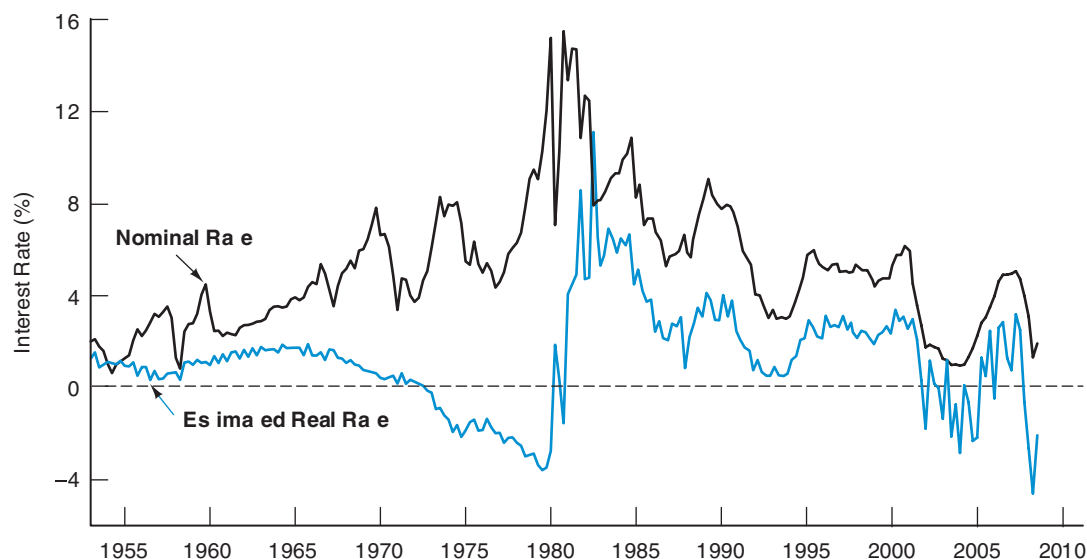


FIG RE 10-1 Real and Nominal Interest Rates (Three-Month Treasury Bill), 1953–2008

Sources: Nominal rates from www.federalreserve.gov/releases/H15. The real rate is constructed using the procedure outlined in Frederic S. Mishkin, “The Real Interest Rate: An Empirical Investigation,” *Carnegie-Rochester Conference Series on Public Policy* 15 (1981): 151–200. These procedures involve estimating expected inflation as a function of past interest rates, inflation, and time trends and then subtracting the expected inflation measure from the nominal interest rate.

which presents estimates from 1953 to 2008 of the real and nominal interest rates on three-month U.S. Treasury bills, shows us that nominal and real rates often do not move together. (This is also true for nominal and real interest rates in Canada and the rest of the world.) By the standard of nominal interest rates, you would have thought that credit market conditions were tight in this period because it was expensive to borrow. However, the estimates of the real rates indicate that you would have been mistaken. In real terms, the cost of borrowing was actually quite low.⁸

⁸ Because most interest income in Canada is subject to income taxes, the true earnings in real terms from holding a debt instrument are not reflected by the real interest rate defined by the Fisher equation but rather by the *after-tax real interest rate*, which equals the nominal interest rate *after income tax payments have been subtracted*, minus the expected inflation rate. For a person facing a 30% tax rate, the after-tax interest rate earned on a bond yielding 10% is only 7% because 30% of the interest income must be paid to the CRA. Thus the after-tax real interest rate on this bond when expected inflation is 5% equals 2% (= 7% – 5%). More generally, the after-tax real interest rate can be expressed as

$$i(1 - \tau) - \pi^e$$

where τ = the income tax rate.

This formula for the after-tax real interest rate also provides a better measure of the effective cost of borrowing for many corporations in Canada because in calculating income taxes, they can deduct interest payments on loans from their income. Thus if you face a 30% tax rate and take out a business loan with a 10% interest rate, you are able to deduct the 10% interest payment and thus lower your business taxes by 30% of this amount. Your after-tax nominal cost of borrowing is then 7% (10% minus 30% of the 10% interest payment), and when the expected inflation rate is 5%, the effective cost of borrowing in real terms is again 2% (= 7% – 5%).

As the example (and the formula) indicates, after-tax real interest rates are always below the real interest rate defined by the Fisher equation. For a further discussion of measures of after-tax real interest rates, see Frederic S. Mishkin, “The Real Interest Rate: An Empirical Investigation,” *Carnegie-Rochester Conference Series on Public Policy* 15 (1981): 151–200.

Formerly, real interest rates in Canada were not observable; only nominal rates were reported. This all changed on December 10, 1991, when the government of Canada began to issue **indexed bonds**, whose interest and principal payments are adjusted for changes in the price level (see the FYI box With Real Return Bonds, Real Interest Rates Have Become Observable in Canada).

FYI

With Real Return Bonds, Real Interest Rates Have Become Observable in Canada

On December 10, 1991, the Canadian government issued coupon bonds whose coupon payment and face value are indexed to the Consumer Price Index (CPI). These securities are known as *Real Return Bonds* and are designed to provide investors with a known real return if held to maturity. Other countries such as the United Kingdom, Australia, and Sweden also issue similar indexed securities, and the U.S. Treasury joined the group (in September 1998) by issuing TIPS (Treasury Inflation Protection Securities).

These indexed securities have successfully acquired a niche in the bond market, enabling governments to raise more funds. In addition, because their interest and principal payments are adjusted for changes in the price level, the interest rate on these bonds provides a direct measure of a real interest rate. These indexed

bonds are very useful to policy makers, especially monetary policy makers, because by subtracting their interest rate from a nominal interest rate on a nonindexed bond, they generate more insight into expected inflation, a valuable piece of information.

For example, on January 28, 2009, the interest rate on long-term Canada bonds was 3.72%, while that on the long-term Real Return Bond was 2.26%. Thus, the implied expected inflation rate, derived from the difference between these two rates, was 1.46%. The private sector finds the information provided by Real Return Bonds very useful: Many financial institutions routinely publish the expected Canadian inflation rate derived from these bonds.

APPLICATION

Calculating the Principal and Coupon Payment of Real Return Bonds

Consider a real return bond with a face value of \$1000 and a coupon yield of 2%. Calculate the principal and coupon payment after one year if the inflation rate is 3%.

Solution

After a year, to account for inflation, the principal will be increased by 3%, from \$1000 to \$1030. The coupon yield is still 2%, but applies to the new principal of \$1030, instead of \$1000. Hence, the coupon payment will be $0.02 \times \$1030 = \20.60 .

SUMMAR

1. The yield to maturity, which is the measure that most accurately reflects the interest rate, is the interest rate that equates the present value of future payments of a debt instrument with its value today. Application of this principle reveals that bond prices and interest rates are negatively related: when the interest rate rises, the price of the bond must fall, and vice versa.
2. The return on a security, which tells you how well you have done by holding this security over a stated period of time, can differ substantially from the interest rate as measured by the yield to maturity. Long-term bond prices have substantial fluctuations when interest rates change and thus bear interest-rate risk. The resulting capital gains and losses can be large, which is why long-term bonds are not considered to be safe assets with a sure return.
3. The real interest rate is defined as the nominal interest rate minus the expected rate of inflation. It is a better measure of the incentives to borrow and lend than the nominal interest rate, and it is a more accurate indicator of the tightness of credit market conditions than the nominal interest rate.

KEY TERMS

cash flows, p. 59	face value (par value), p. 62	present value, p. 59
consol (perpetuity), p. 67	fixed-payment loan (fully amortized loan), p. 62	rate of capital gain, p. 72
coupon bond, p. 62	indexed bond, p. 79	real interest rate, p. 76
coupon rate, p. 62	interest-rate risk, p. 74	real terms, p. 77
current yield, p. 68	nominal interest rate, p. 76	return (rate of return), p. 71
discount bond (zero-coupon bond), p. 62	present discounted value, p. 59	simple loan, p. 59
		yield to maturity, p. 63

QUESTIONS

You will find the answers to the questions marked with an asterisk in the Textbook Resources section of your MyEconLab.

- *1. Write down the formula that is used to calculate the yield to maturity on a 20-year 10% coupon bond with \$1000 face value that sells for \$2000.
2. If there is a decline in interest rates, which would you rather be holding, long-term bonds or short-

term bonds? Why? Which type of bond has the greater interest-rate risk?

- *3. Francine the Financial Adviser has just given you the following advice: "Long-term bonds are a great investment because their interest rate is over 20%." Is Francine necessarily right?

QUANTITATIVE PROBLEMS

- *1. Would a dollar tomorrow be worth more to you today when the interest rate is 20% or when it is 10%?
2. You have just won \$20 million in a provincial lottery, which promises to pay you \$1 million (tax-free) every year for the next 20 years. Have you really won \$20 million?
- *3. If the interest rate is 10%, what is the present value of a security that pays you \$1100 next year, \$1210 the year after, and \$1331 the year after that?
4. If the security in Problem 3 sold for \$4000, is the yield to maturity greater or less than 10%? Why?
5. What is the yield to maturity on a \$1000 face-value discount bond maturing in one year that sells for \$800?
- *6. What is the yield to maturity on a simple loan for \$1 million that requires a repayment of \$2 million in five years' time?
7. To pay for university, you have just taken out a \$1000 government loan that makes you pay \$126 per year for 25 years. However, you don't have to start making these payments until you graduate from university two years from now. Why is the yield to maturity necessarily less than 12%, the yield to maturity on a normal \$1000 fixed-payment loan in which you pay \$126 per year for 25 years?
- *8. Which \$1000 bond has the higher yield to maturity, a 20-year bond selling for \$800 with a current yield of 15% or a one-year bond selling for \$800 with a current yield of 5%?

9. Pick five Canada bonds from the bond page of the newspaper, and calculate the current yield. Note when the current yield is a good approximation of the yield to maturity.
- *10. You are offered two bonds, a one-year Canada bond with a yield to maturity of 9% and a one-year treasury bill with a yield on a discount basis of 8.9%. Which would you rather own?
11. If mortgage rates rise from 5% to 10% but the expected rate of increase in housing prices rises from 2% to 9%, are people more or less likely to buy houses?
- *12. Interest rates were lower in the mid-1980s than they were in the late 1970s, yet many economists have commented that real interest rates were actually much higher in the mid-1980s than in the late 1970s. Does this make sense? Do you think that these economists are right?
13. You borrowed \$1000 on January 1 and must repay a total amount of \$1060 exactly a year later.
 - a. What is the interest paid?
 - b. What is the interest rate?
- *14. Consider a perpetuity that has a coupon of \$100 per year.
 - a. What is the price of the perpetuity if the yield to maturity is 5%?
 - b. If the yield to maturity doubles, what will happen to the price?
- *15. Suppose that the interest rate is 5%. Which of the following statements are true and which are false?
 - a. \$57 today is equivalent to \$61 one year from now.
 - b. \$5000 today is equivalent to \$5250 one year from now.
 - c. \$37.80 one year from now is equivalent to \$36 today.

CANSIM Questions

16. Get the quarterly data from 1953 to 2009 on the three-month T-bill rate (CANSIM series V122541) and the total consumer price index (series V41690973) from the Textbook Resources area of the MyEconLab.
 - a. Calculate the (actual) annual inflation rate, using the formula

$$\pi = 4 \times 100 \times (P_{+1} - P)/P$$
 - b. Plot the nominal interest rate, i , and the inflation rate, π .
 - c. Assume that the expected inflation rate is the same as the actual inflation rate (a restrictive assumption!) and calculate the real interest rate, i_r .
 - d. Plot the nominal and real interest rates on a graph.
 - e. What is the relationship between the nominal interest rate, i , and the real interest rate, i_r , over this period?
17. Get the monthly data from 1991 to 2009 for the interest rate on long-term Canada Real Return Bonds (CANSIM series V122553) from the Textbook Resources area of the MyEconLab.
 - a. Plot this real interest rate, i_r .
 - b. Has the real interest rate been rising or falling over the sample period?
 - c. What is the mean real interest rate over the sample period? What is its standard deviation (the standard deviation is the square root of the variance)?

WEB EXERCISES

1. Investigate the data on interest rates available from the Bank of Canada at www.bankofcanada.ca. Answer the following questions.
 - a. What is the difference in the interest rates on 10-year and 2-year bonds?
 - b. What is the difference in the interest rate on long-term government of Canada bonds and Real Return Bonds?
 - c. What is the difference in the interest rate on long-term Government of Canada bonds and corporate bonds?
2. Figure 4-1 (page 78) shows the estimated real and nominal rates for three-month U.S. treasury bills. Go to www.martincapital.com/main/charts.html and click on the relevant link under "Charts of Interest Rates and Yields."
 - a. Compare the three-month real rate to the long-term real rate. Which is greater?
 - b. Compare the short-term nominal rate to the long-term nominal rate. Which appears most volatile?



myeconlab

Be sure to visit the MyEconLab website at www.myeconlab.com. This online homework and tutorial system puts you in control of your own learning with study and practice tools directly correlated to this chapter content.

On the MyEconLab website you will find the following appendix and mini-case for this chapter:

Appendix 4.1: Measuring Interest Rate Risk: Duration

Mini-Case 4.1: Interest Rates, Bond Yields, and Duration

CHAPTER 11

The Behaviour of Interest Rates

LEARNING OBJECTIVES

After studying this chapter you should be able to

1. describe how the demand and supply analysis for bonds provides one theory of how nominal interest rates are determined
2. explain how the demand and supply analysis for money, known as the liquidity preference framework, provides an alternative theory of interest-rate determination
3. outline the factors that cause interest rates to change
4. characterize the effects of monetary policy on interest rates: the liquidity effect, the income effect, the price-level effect, and the expected-inflation effect

PREVIEW

In the early 1950s, nominal interest rates on three-month treasury bills were about 1% at an annual rate; by 1981, they had reached over 20%; in the early 2000s and in 2008 they fell below 2%. What explains these substantial fluctuations in interest rates? One reason why we study money, banking, and financial markets is to provide some answers to this question.

In this chapter we examine how the overall level of *nominal* interest rates (which we refer to as simply “interest rates”) is determined and what factors influence their behaviour. We learned in Chapter 4 that interest rates are negatively related to the price of bonds, so if we can explain why bond prices change, we can also explain why interest rates fluctuate. We make use of supply and demand analysis for markets for bonds and money to examine how interest rates change.

In order to derive a demand curve for assets like money or bonds, the first step in our analysis, we must first understand what determines the demand for these assets. We do this by developing an economic theory known as the *theory of asset demand*, which outlines criteria that are important when deciding how much of an asset to buy. Armed with this theory, we can then go on to derive the demand curve for bonds or money. After deriving supply curves for these assets, we develop the concept of market equilibrium, the point at which the quantity supplied equals the quantity demanded. Then we use this model to explain changes in equilibrium interest rates.

Because interest rates on different securities tend to move together, in this chapter we will act as if there is only one type of security and one interest rate in the entire economy. In the following chapter, we expand our analysis to look at why interest rates on different types of securities differ.

DETERMINANTS OF ASSET DEMAND

Before going on to our supply and demand analysis of the bond market and the market for money, we must first understand what determines the quantity demanded of an asset. Recall that an asset is a piece of property that is a store of value. Items such as money, bonds, stocks, art, land, houses, farm equipment, and manufacturing machinery are all assets. Facing the question of whether to buy and hold an asset or whether to buy one asset rather than another, an individual must consider the following factors:

1. **Wealth**, the total resources owned by the individual, including all assets
2. **Expected return** (the return expected over the next period) on one asset relative to alternative assets
3. **Risk** (the degree of uncertainty associated with the return) on one asset relative to alternative assets
4. **Liquidity** (the ease and speed with which an asset can be turned into cash) relative to alternative assets.

Wealth

When we find that our wealth has increased, we have more resources available with which to purchase assets, and so, not surprisingly, the quantity of assets we demand increases. Therefore, the effect of changes in wealth on the quantity demanded of an asset can be summarized as follows: ***holding everything else constant, an increase in wealth raises the quantity demanded of an asset.***

Expected Return

In Chapter 4 we saw that the return on an asset (such as a bond) measures how much we gain from holding that asset. When we make a decision to buy an asset, we are influenced by what we expect the return on that asset to be. If a Bell Canada bond, for example, has a return of 15% half the time and 5% the other half of the time, its expected return (which you can think of as the average return) is 10% ($= 0.5 \times 15\% + 0.5 \times 5\%$).¹ If the expected return on the Bell bond rises relative to expected returns on alternative assets, holding everything else constant, then it becomes more desirable to purchase it, and the quantity demanded increases. This can occur in either of two ways: (1) when the expected return on the Bell bond rises while the return on an alternative asset—say, stock in TD Canada Trust—remains unchanged or (2) when the return on the alternative asset, the TD Canada Trust stock, falls while the return on the Bell bond remains unchanged. To summarize, ***an increase in an asset's expected return relative to that of an alternative asset, holding everything else unchanged, raises the quantity demanded of the asset.***

Risk

The degree of risk or uncertainty of an asset's returns also affects the demand for the asset. Consider two assets, stock in Fly-by-Night Airlines and stock in Feet-on-the-Ground Bus Company. Suppose that Fly-by-Night stock has a return of 15% half the time and 5% the other half of the time, making its expected return

¹ If you are interested in finding out more information on how to calculate expected returns, as well as standard deviations of returns that measure risk, you can look at an appendix to this chapter that describes models of asset pricing on this book's MyEconLab at www.pearsoned.ca/myeconlab. This appendix also describes how diversification lowers the overall risk of a portfolio and has a discussion of systematic risk and basic asset pricing models such as the capital asset pricing model and arbitrage pricing theory.

10%, while stock in Feet-on-the-Ground has a fixed return of 10%. Fly-by-Night stock has uncertainty associated with its returns and so has greater risk than stock in Feet-on-the-Ground, whose return is a sure thing.

A *risk-averse* person prefers stock in Feet-on-the-Ground (the sure thing) to Fly-by-Night stock (the riskier asset), even though the stocks have the same expected return, 10%. By contrast, a person who prefers risk is a *risk-preferer* or *risk-lover*. Most people are risk-averse, especially in their financial decisions: everything else being equal, they prefer to hold the less-risky asset. Hence, **holding everything else constant, if an asset's risk rises relative to that of alternative assets, its quantity demanded will fall.**

Liquidity

Another factor that affects the demand for an asset is how quickly it can be converted into cash at low cost—its liquidity. An asset is liquid if the market in which it is traded has depth and breadth, that is, if the market has many buyers and sellers. A house is not a very liquid asset because it may be hard to find a buyer quickly; if a house must be sold to pay off bills, it might have to be sold for a much lower price. And the transaction costs in selling a house (broker's commissions, lawyer's fees, and so on) are substantial. A Canadian government treasury bill, by contrast, is a highly liquid asset. It can be sold in a well-organized market where there are many buyers, so it can be sold quickly at low cost. **The more liquid an asset is relative to alternative assets, holding everything else unchanged, the more desirable it is, and the greater will be the quantity demanded.**

Theory of Asset Demand

All the determining factors we have just discussed can be assembled into the **theory of asset demand**, which states that, holding all of the other factors constant:

1. The quantity demanded of an asset is positively related to wealth.
2. The quantity demanded of an asset is positively related to its expected return relative to alternative assets.
3. The quantity demanded of an asset is negatively related to the risk of its returns relative to alternative assets.
4. The quantity demanded of an asset is positively related to its liquidity relative to alternative assets.

These results are summarized in Table 5-1.

TABLE 11-1 Response of the Quantity of an Asset Demanded to Changes in Income or Wealth, Expected Returns, Risk, and Liquidity

Variable	Change in Variable	Change in Quantity Demanded
Wealth	↑	↑
Expected return relative to other assets	↑	↑
Risk relative to other assets	↑	↓
Liquidity relative to other assets	↑	↑

Note: Only increases (↑) in the variables are shown. The effect of decreases in the variables on the change in demand would be the opposite of those indicated in the rightmost column.

SUPPLY AND DEMAND IN THE BOND MARKET

Our first approach to the analysis of interest-rate determination looks at supply and demand in the bond market to see how the price of bonds is determined. With our understanding of how interest rates are measured from the previous chapter, we then recognize that each bond price is associated with a particular level of interest rates. Specifically, the negative relationship between bond prices and interest rates means that when we see that the bond price rises, the interest rate falls (or vice versa).

The first step in the analysis is to obtain a bond **demand curve**, which shows the relationship between the quantity demanded and the price when all other economic variables are held constant (that is, values of other variables are taken as given). You may recall from previous economics courses that the assumption that all other economic variables are held constant is called *ceteris paribus*, which means “other things being equal” in Latin.

Demand Curve

To clarify our analysis, let us consider the demand for one-year discount bonds, which make no coupon payments but pay the owner the \$1000 face value in a year. If the holding period is one year, then, as we saw in Chapter 4, the return on the bonds is known absolutely and is equal to the interest rate as measured by the yield to maturity. This means that the expected return on this bond is equal to the interest rate i , which, using Equation 6 from Chapter 4 (page 70), is

$$i = RET^e = \frac{F - P}{P}$$

where

i = interest rate = yield to maturity
 RET^e = expected return
 F = face value of the discount bond
 P = initial purchase price of the discount bond

This formula shows that a particular value of the interest rate corresponds to each bond price. If the bond sells for \$950, the interest rate and expected return is

$$\frac{\$1000 - \$950}{\$950} = 0.053 = 5.3\%$$

At this 5.3% interest rate and expected return corresponding to a bond price of \$950, let us assume that the quantity of bonds demanded is \$100 billion, which is plotted as point A in Figure 5-1.

At a price of \$900, the interest rate and expected return are

$$\frac{\$1000 - \$900}{\$900} = 0.111 = 11.1\%$$

Because the expected return on these bonds is higher, with all other economic variables (such as income, expected returns on other assets, risk, and liquidity) held constant, the quantity demanded of bonds will be higher as predicted by the theory of asset demand. Point B in Figure 5-1 shows that the quantity of bonds demanded at the price of \$900 has risen to \$200 billion. Continuing with this reasoning, if the bond price is \$850 (interest rate and expected return = 17.6%), the quantity of bonds demanded (point C) will be greater than at point B. Similarly, at

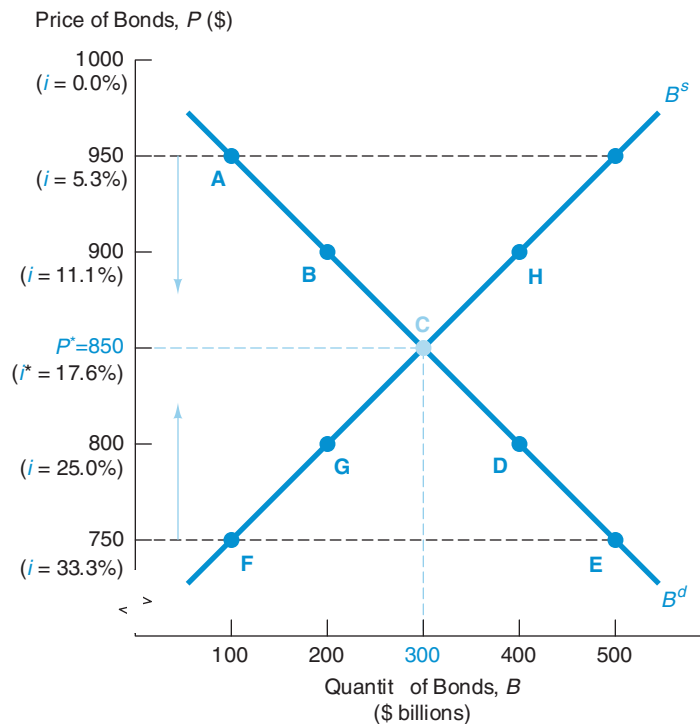


FIG RE 11-1 Supply and Demand for Bonds

Equilibrium in the bond market occurs at point C, the intersection of the demand curve B^d and the bond supply curve B^s . The equilibrium price is $P^* = \$850$, and the equilibrium interest rate is $i^* = 17.6\%$. (Note: P and i increase in opposite directions.)

the lower prices of \$800 (interest rate = 25%) and \$750 (interest rate = 33.3%), the quantity of bonds demanded will be even higher (points D and E). The curve B^d , which connects these points, is the demand curve for bonds. It has the usual downward slope, indicating that at lower prices of the bond (everything else being equal), the quantity demanded is higher.²

Supply Curve

An important assumption behind the demand curve for bonds in Figure 5-1 is that all other economic variables besides the bond's price and interest rate are held constant. We use the same assumption in deriving a **supply curve**, which shows the relationship between the quantity supplied and the price when all other economic variables are held constant.

When the price of the bonds is \$750 (interest rate = 33.3%), point F shows that the quantity of bonds supplied is \$100 billion for the example we are considering. If the price is \$800, the interest rate is the lower rate of 25%. Because at this interest rate it is now less costly to borrow by issuing bonds, firms will be willing to borrow more through bond issues, and the quantity of bonds supplied is at the higher level of \$200 billion (point G). An even higher price of \$850, corresponding to a lower interest rate of 17.6%, results in a larger quantity of bonds supplied

² Note that although our analysis indicates that the demand curve is downward-sloping, it does not imply that the curve is a straight line. For ease of exposition, however, we will draw demand curves and supply curves as straight lines.

of \$300 billion (point C). Higher prices of \$900 and \$950 result in even greater quantities of bonds supplied (points H and I). The B^s curve, which connects these points, is the supply curve for bonds. It has the usual upward slope found in supply curves, indicating that as the price increases (everything else being equal), the quantity supplied increases.

Market Equilibrium

In economics, **market equilibrium** occurs when the amount that people are willing to buy (*demand*) equals the amount that people are willing to sell (*supply*) at a given price. In the bond market, this is achieved when the quantity of bonds demanded equals the quantity of bonds supplied:

$$B^d = B^s \quad (1)$$

In Figure 11-1, equilibrium occurs at point C, where the demand and supply curves intersect at a bond price of \$850 (interest rate of 17.6%) and a quantity of bonds of \$300 billion. The price of $P^* = \$850$, where the quantity demanded equals the quantity supplied, is called the *equilibrium* or *market-clearing* price. Similarly, the interest rate of $i^* = 17.6\%$ that corresponds to this price is called the *equilibrium* or *market-clearing* interest rate.

The concepts of market equilibrium and equilibrium price or interest rate are useful because there is a tendency for the market to head toward them. We can see that it does in Figure 5-1 by first looking at what happens when we have a bond price that is above the equilibrium price. When the price of bonds is set too high, at, say, \$950, the quantity of bonds supplied at point I is greater than the quantity of bonds demanded at point A. A situation like this, in which the quantity of bonds supplied exceeds the quantity of bonds demanded, is called a condition of **excess supply**. Because people want to sell more bonds than others want to buy, the price of the bonds will fall, and this is why the downward arrow is drawn in the figure at the bond price of \$950. As long as the bond price remains above the equilibrium price, there will continue to be an excess supply of bonds, and the price will continue to fall. This will stop only when the price has reached the equilibrium price of \$850, where the excess supply of bonds will be eliminated.

Now let's look at what happens when the price of bonds is below the equilibrium price. If the price of the bonds is set too low, say at \$750, the quantity demanded at point E is greater than the quantity supplied at point F. This is called a condition of **excess demand**. People now want to buy more bonds than others are willing to sell, and so the price of bonds will be driven up. This is illustrated by the upward arrow drawn in the figure at the bond price of \$750. Only when the excess demand for bonds is eliminated by the price rising to the equilibrium level of \$850 is there no further tendency for the price to rise.

We can see that the concept of equilibrium price is a useful one because it indicates where the market will settle. Because each price on the vertical axis of Figure 5-1 shows a corresponding interest rate value, the same diagram also shows that the interest rate will head toward the equilibrium interest rate of 17.6%. When the interest rate is below the equilibrium interest rate, as it is when it is at 5.3%, the price of the bond is above the equilibrium price, and there will be an excess supply of bonds. The price of the bond then falls, leading to a rise in the interest rate toward the equilibrium level. Similarly, when the interest rate is above the equilibrium level, as it is when it is at 33.3%, there is excess demand for bonds, and the bond price will rise, driving the interest rate back down to the equilibrium level of 17.6%.

Supply and Demand Analysis

Our Figure 5-1 is a conventional supply and demand diagram with price on the left vertical axis and quantity on the horizontal axis. Because the interest rate that corresponds to each bond price is also marked on the vertical axis, this diagram allows us to read the equilibrium interest rate, giving us a model that describes the determination of interest rates. It is important to recognize that a supply and demand diagram like Figure 5-1 can be drawn for *any* type of bond because the interest rate and price of a bond are *always* negatively related for any type of bond, whether a discount bond or a coupon bond.

An important feature of the analysis here is that supply and demand are always in terms of *stocks* (amounts at a given point in time) of assets, not in terms of *flows*. The **asset market approach** for understanding behaviour in financial markets—which emphasizes stocks of assets rather than flows in determining asset prices—is now the dominant methodology used by economists because correctly conducting analyses in terms of flows is very tricky, especially when we encounter inflation.³

CHANGES IN EQUILIBRIUM INTEREST RATES

We will now use the supply and demand framework for bonds to analyze why interest rates change. To avoid confusion, it is important to make the distinction between *movements along* a demand (or supply) curve and *shifts in* a demand (or supply) curve. When quantity demanded (or supplied) changes as a result of a change in the price of the bond (or, equivalently, a change in the interest rate), we have a *movement along* the demand (or supply) curve. The change in the quantity demanded when we move from point A to B to C in Figure 5-1, for example, is a movement along a demand curve. A *shift in* the demand (or supply) curve, by contrast, occurs when the quantity demanded (or supplied) changes *at each given price (or interest rate)* of the bond in response to a change in some other factor besides the bond's price or interest rate. When one of these factors changes, causing a shift in the demand or supply curve, there will be a new equilibrium value for the interest rate.

In the following pages we will look at how the supply and demand curves shift in response to changes in variables, such as expected inflation and wealth, and what effects these changes have on the equilibrium value of interest rates.

Shifts in the Demand for Bonds

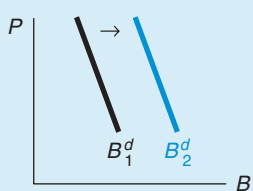
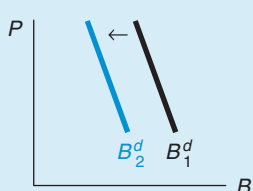
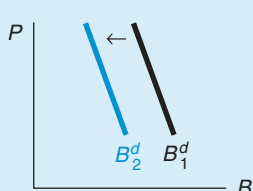
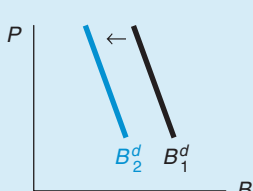
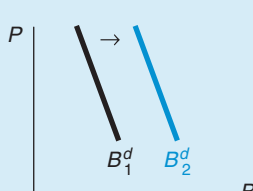
The theory of asset demand developed at the beginning of the chapter provides a framework for deciding what factors cause the demand curve for bonds to shift. These factors include changes in four parameters:

1. Wealth
2. Expected returns on bonds relative to alternative assets
3. Risk of bonds relative to alternative assets
4. Liquidity of bonds relative to alternative assets

³ The asset market approach developed in the text is useful in understanding not only how interest rates behave but also how any asset price is determined. A second appendix to this chapter, which is on this book's MyEconLab at www.pearsoned.ca/m_econlab, shows how the asset market approach can be applied to understanding the behaviour of commodity markets, in particular, the gold market. The analysis of the bond market that we have developed here has another interpretation using a different terminology and framework involving the supply and demand for loanable funds. This loanable funds framework is discussed in a third appendix to this chapter, which is also on MyEconLab.

To see how a change in each of these factors (holding all other factors constant) can shift the demand curve, let us look at some examples. (As a study aid, Table 5-2 summarizes the effects of changes in these factors on the bond demand curve.)

TABLE 11-2 Factors That Shift the Demand Curve for Bonds

Variable	Change in Variable	Change in Quantity Demanded at Each Bond Price	Shift in Demand Curve
Wealth	↑	↑	
Expected interest rate	↑	↓	
Expected inflation	↑	↓	
Riskiness of bonds relative to other assets	↑	↓	
Liquidity of bonds relative to other assets	↑	↑	

Note: Only increases (↑) in the variables are shown. The effect of decreases in the variables on the change in demand would be the opposite of those indicated in the remaining columns.

WEALTH When the economy is growing rapidly in a business cycle expansion and wealth is increasing, the quantity of bonds demanded at each bond price (or interest rate) increases, as shown in Figure 5-2. To see how this works, consider point B on the initial demand curve for bonds B_1^d . With higher wealth, the quantity of bonds demanded at the same price must rise, to point B'. Similarly, the higher wealth causes the quantity demanded at the same bond price to rise to point D'. Continuing with this reasoning for every point on the initial demand curve B_1^d , we can see that the demand curve shifts to the right from B_1^d to B_2^d as is indicated by the arrows.

The conclusion we have reached is that ***in a business cycle expansion with growing wealth, the demand for bonds rises and the demand curve for bonds shifts to the right***. Using the same reasoning, ***in a recession, when income and wealth are falling, the demand for bonds falls, and the demand curve shifts to the left***.

Another factor that affects wealth is the public's propensity to save. If households save more, wealth increases and, as we have seen, the demand for bonds rises and the demand curve for bonds shifts to the right. Conversely, if people save less, wealth and the demand for bonds will fall and the demand curve shifts to the left.

EXPECTED RETURNS For a one-year discount bond and a one-year holding period, the expected return and the interest rate are identical, so nothing besides today's interest rate affects the expected return.

For bonds with maturities of greater than one year, the expected return may differ from the interest rate. For example, we saw in Chapter 4, Table 4-2 (page 74), that a rise in the interest rate on a long-term bond from 10 to 20% would lead to a sharp decline in price and a very negative return. Hence, if people begin to think that interest rates will be higher next year than they had originally anticipated, the

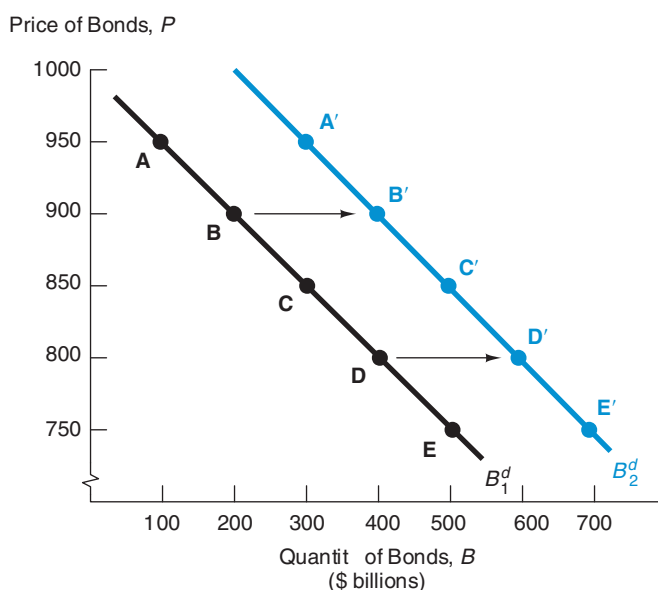


FIG RE 11-2 Shift in the Demand Curve for Bonds

When the demand for bonds increases, the demand curve shifts to the right as shown.

expected return today on long-term bonds will fall, and the quantity demanded will fall at each interest rate. **Higher expected interest rates in the future lower the expected return for long-term bonds, decrease the demand, and shift the demand curve to the left.**

By contrast, a revision downward of expectations of future interest rates would mean that long-term bond prices would be expected to rise more than originally anticipated, and the resulting higher expected return today would raise the quantity demanded at each bond price and interest rate. **Lower expected interest rates in the future increase the demand for long-term bonds and shift the demand curve to the right** (as in Figure 5-2).

Changes in expected returns on other assets can also shift the demand curve for bonds. If people suddenly became more optimistic about the stock market and began to expect higher stock prices in the future, both expected capital gains and expected returns on stocks would rise. With the expected return on bonds held constant, the expected return on bonds today relative to stocks would fall, lowering the demand for bonds and shifting the demand curve to the left.

A change in expected inflation is likely to alter expected returns on physical assets (also called *real assets*) such as automobiles and houses, which affects the demand for bonds. An increase in expected inflation, say, from 5% to 10%, will lead to higher prices on cars and houses in the future and hence higher nominal capital gains. The resulting rise in the expected returns today on these real assets will lead to a fall in the expected return on bonds relative to the expected return on real assets today and thus cause the demand for bonds to fall. Alternatively, we can think of the rise in expected inflation as lowering the real interest rate on bonds, and the resulting decline in the relative expected return on bonds causes the demand for bonds to fall. **An increase in the expected rate of inflation lowers the expected return for bonds, causing their demand to decline and the demand curve to shift to the left.**

RISK If prices in the bond market become more volatile, the risk associated with bonds increases, and bonds become a less attractive asset. **An increase in the riskiness of bonds causes the demand for bonds to fall and the demand curve to shift to the left.**

Conversely, an increase in the volatility of prices in another asset market, such as the stock market, would make bonds more attractive. **An increase in the riskiness of alternative assets causes the demand for bonds to rise and the demand curve to shift to the right** (as in Figure 5-2).

LIQUIDITY If more people started trading in the bond market and as a result it became easier to sell bonds quickly, the increase in their liquidity would cause the quantity of bonds demanded at each interest rate to rise. **Increased liquidity of bonds results in an increased demand for bonds, and the demand curve shifts to the right** (see Figure 5-2). **Similarly, increased liquidity of alternative assets lowers the demand for bonds and shifts the demand curve to the left.** The reduction of brokerage commissions for trading common stocks that occurred when the fixed-rate commission structure was abolished in 1975, for example, increased the liquidity of stocks relative to bonds, and the resulting lower demand for bonds shifted the demand curve to the left.

Shifts in the Supply of Bonds

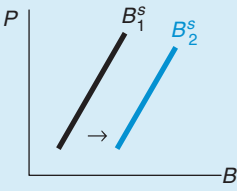
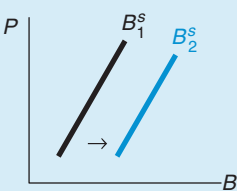
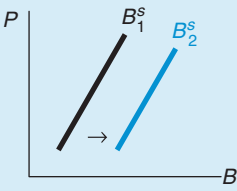
Certain factors can cause the supply curve for bonds to shift, among them:

- 1. Expected profitability of investment opportunities
- 2. Expected inflation
- 3. Government activities

We will look at how the supply curve shifts when each of these factors changes (all others remaining constant). (As a study aid, Table 5-3 summarizes the effects of changes in these factors on the bond supply curve.)

EXPECTED PROFITABILITY OF INVESTMENT OPPORTUNITIES The more profitable plant and equipment investments that a firm expects it can make, the more willing it will be to borrow in order to finance these investments. When the economy is growing rapidly, as in a business cycle expansion, investment opportunities that are expected to be profitable abound, and the quantity of bonds supplied at any given bond price will increase (see Figure 5-3). *Therefore, in a business cycle expansion, the supply of bonds increases, and the supply curve shifts to the right. Likewise, in a recession, when there are far fewer expected profitable invest-*

TABLE 11-3 Factors That Shift the Supply Curve of Bonds

Variable	Change in Variable	Change in Quantity Supplied at Each Bond Price	Shift in Supply Curve
Profitability of investments	↑	↑	
Expected inflation	↑	↑	
Government deficit	↑	↑	

Note: Only increases (↑) in the variables are shown. The effect of decreases in the variables on the change in supply would be the opposite of those indicated in the remaining columns.

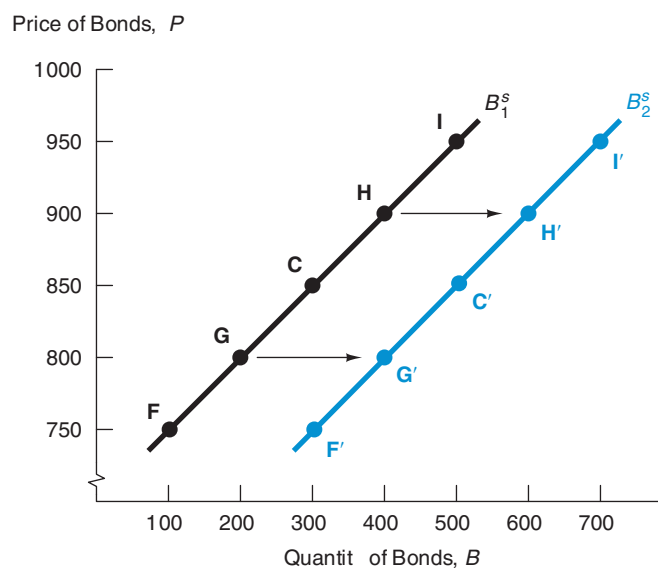


FIG RE 11-3 Shift in the Supply Curve for Bonds

When the supply of bonds increases, the supply curve shifts to the right.

ment opportunities, the supply of bonds falls and the supply curve shifts to the left.

EXPECTED INFLATION As we saw in Chapter 4, the real cost of borrowing is more accurately measured by the real interest rate, which equals the (nominal) interest rate minus the expected inflation rate. For a given interest rate (and bond price), when expected inflation increases, the real cost of borrowing falls; hence the quantity of bonds supplied increases at any given bond price. **An increase in expected inflation causes the supply of bonds to increase and the supply curve to shift to the right** (see Figure 5-3).

GOVERNMENT BUDGET The activities of the government can influence the supply of bonds in several ways. The Canadian government issues bonds to finance government deficits, the gap between the government's expenditures and its revenues. When these deficits are large, the government sells more bonds, and the quantity of bonds supplied at each bond price increases. **Higher government deficits increase the supply of bonds and shift the supply curve to the right** (see Figure 5-3). **On the other hand, government surpluses, as have occurred in recent years, decrease the supply of bonds and shift the supply curve to the left.**

Provincial and municipal governments and other government agencies also issue bonds to finance their expenditures, and this can also affect the supply of bonds. We will see in later chapters that the conduct of monetary policy involves the purchase and sale of bonds, which in turn influences the supply of bonds.

We now can use our knowledge of how supply and demand curves shift to analyze how the equilibrium interest rate can change. The best way to do this is to pursue several applications that are particularly relevant to our understanding of how monetary policy affects interest rates. In going through these applications, keep two things in mind:

1. When you examine the effect of a variable change, remember that we are assuming that all other variables are unchanged; that is, we are making use of the *ceteris paribus* assumption.
2. Remember that the interest rate is negatively related to the bond price, so when the equilibrium bond price rises, the equilibrium interest rate falls. Conversely, if the equilibrium bond price moves downward, the equilibrium interest rate rises.

APPLICATION

Changes in the Interest Rate Due to Expected Inflation: The Fisher Effect

We have already done most of the work to evaluate how a change in expected inflation affects the nominal interest rate in that we have already analyzed how a change in expected inflation shifts the supply and demand curves. Figure 5-4 shows the effect on the equilibrium interest rate of an increase in expected inflation.

Suppose that expected inflation is initially 5% and the initial supply and demand curves B_1^s and B_1^d intersect at point 1, where the equilibrium bond price is P_1 . If expected inflation rises to 10%, the expected return on bonds relative to real assets falls for any given bond price and interest rate. As a result, the demand for bonds falls, and the demand curve shifts to the left from B_1^d to B_2^d . The rise in expected inflation also shifts the supply curve. At any given bond price and interest rate, the real cost of borrowing has declined, causing the quantity of bonds supplied to increase, and the supply curve shifts to the right, from B_1^s to B_2^s .

When the demand and supply curves shift in response to the change in expected inflation, the equilibrium moves from point 1 to point 2, the intersection of B_2^d and B_2^s . The equilibrium bond price has fallen from P_1 to P_2 , and because the bond price is negatively related to the interest rate, this means that the interest rate has risen. Note that Figure 5-4 has been drawn so that the equilibrium

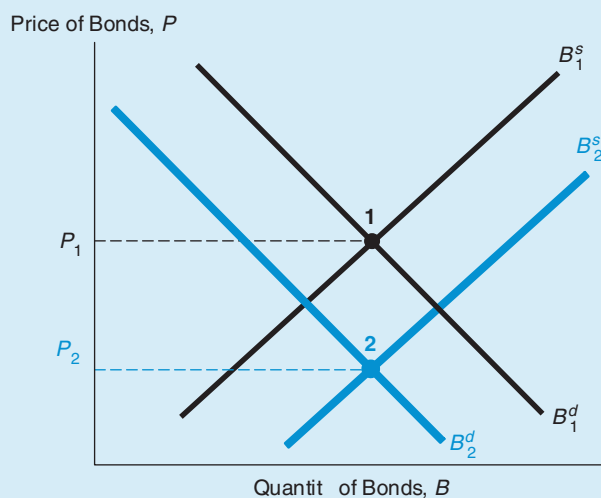


FIG RE 11-4 Response to a Change in Expected Inflation

When expected inflation rises, the supply curve shifts from B_1^s to B_2^s , and the demand curve shifts from B_1^d to B_2^d . The equilibrium moves from point 1 to point 2, with the result that the equilibrium bond price falls from P_1 to P_2 and the equilibrium interest rate rises.

quantity of bonds remains the same for both point 1 and point 2. However, depending on the size of the shifts in the supply and demand curves, the equilibrium quantity of bonds could either rise or fall when expected inflation rises.

Our supply and demand analysis has led us to an important observation: **when expected inflation rises, interest rates will rise.** This result has been named the **Fisher effect**, after Irving Fisher, the economist who first pointed out the relationship of expected inflation to interest rates. The accuracy of this prediction is shown in Figure 5-5 for the United States; a similar figure exists for Canada. The interest rate on three-month U.S. Treasury bills has usually moved along with the expected inflation rate. Consequently, it is understandable that many economists recommend that inflation must be kept low if we want to keep interest rates low.

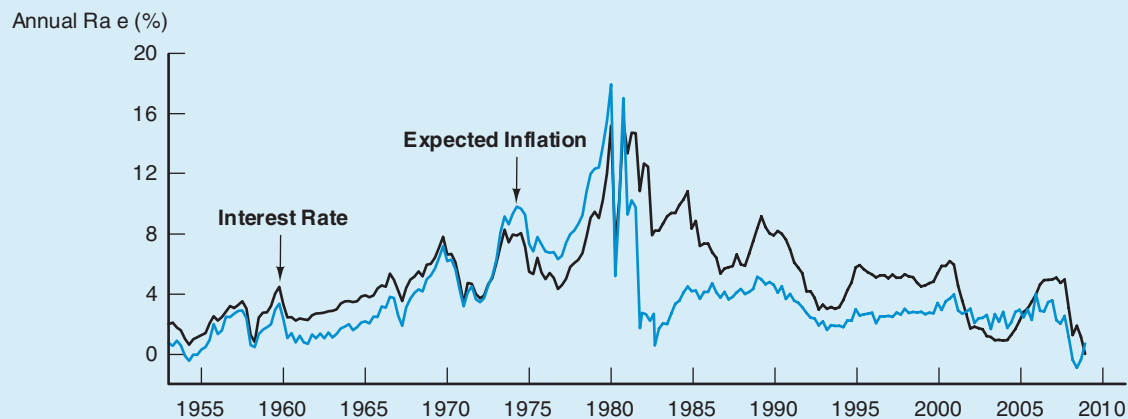


FIG RE 5-5 Expected Inflation and Interest Rates (Three-Month Treasury Bills), 1953–2008

Source: Expected inflation calculated using procedures outlined in Frederic S. Mishkin, "The Real Interest Rate: An Empirical Investigation," *Carnegie-Rochester Conference Series on Public Policy* 15 (1981): 151–200. These procedures involve estimating expected inflation as a function of past interest rates, inflation, and time trends.

APPLICATION

Changes in the Interest Rate Due to a Business Cycle Expansion

Figure 5-6 analyzes the effects of a business cycle expansion on interest rates. In a business cycle expansion, the amount of goods and services being produced in the economy rises, so national income increases. When this occurs, businesses will be more willing to borrow because they are likely to have many profitable investment opportunities for which they need financing. Hence at a given bond price, the quantity of bonds that firms want to sell (that is, the supply of bonds) will increase. This means that in a business cycle expansion, the supply curve for bonds shifts to the right (see Figure 5-6) from B_1^s to B_2^s .

Expansion in the economy will also affect the demand for bonds. As the business cycle expands, wealth is likely to increase, and the theory of asset demand tells

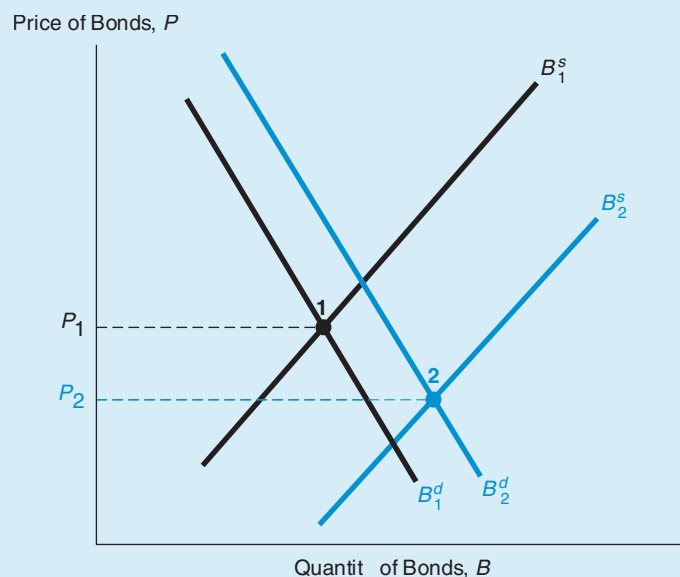


FIG RE 11-6 Response to a Business Cycle Expansion

In a business cycle expansion, when income and wealth are rising, the demand curve shifts rightward from B_1^d to B_2^d , and the supply curve shifts rightward from B_1^s to B_2^s . If the supply curve shifts to the right more than the demand curve, as in this figure, the equilibrium bond price moves down from P_1 to P_2 , and the equilibrium interest rate rises.

us that the demand for bonds will rise as well. We see this in Figure 5-6, where the demand curve has shifted to the right, from B_1^d to B_2^d .

Given that both the supply and demand curves have shifted to the right, we know that the new equilibrium reached at the intersection of B_2^d and B_2^s must also move to the right. However, depending on whether the supply curve shifts more than the demand curve or vice versa, the new equilibrium interest rate can either rise or fall.

The supply and demand analysis used here gives us an ambiguous answer to the question of what will happen to interest rates in a business cycle expansion. The figure has been drawn so that the shift in the supply curve is greater than the shift in the demand curve, causing the equilibrium bond price to fall to P_2 , leading to a rise in the equilibrium interest rate. The reason the figure has been drawn so that a business cycle expansion and a rise in income lead to a higher interest rate is that this is the outcome we actually see in the data. Figure 5-7 plots the movement of the interest rate on three-month treasury bills from 1962 to 2008 and indicates when the business cycle is undergoing recessions (shaded areas). As you can see, the interest rate tends to rise during business cycle expansions and fall during recessions, which is what the supply and demand diagram indicates.



FIG RE 11-7 Business Cycles and Interest Rates (Three-Month Treasury Bills), 1962–2008

Shaded areas indicate periods of recession. The figure shows that interest rates rise during business cycle expansions and fall during contractions, which is what Figure 5-6 suggests would happen.

Source: Statistics Canada CANSIM II Series V122531.

APPLICATION

Explaining Low Japanese Interest Rates

In the 1990s and early 2000s, Japanese interest rates became the lowest in the world. Indeed, in November 1998, an extraordinary event occurred: interest rates on Japanese six-month treasury bills turned slightly negative (see Chapter 4). Why did Japanese rates drop to such low levels?

In the late 1990s, Japan experienced a prolonged recession, which was accompanied by deflation, a negative inflation rate. Using these facts, analysis similar to that used in the preceding application explains the low Japanese interest rates.

Negative inflation caused the demand for bonds to rise because the expected return on real assets fell, thereby raising the relative expected return on bonds and in turn causing the demand curve to shift to the right. The negative inflation also raised the real interest rate and therefore the real cost of borrowing for a given nominal rate, thereby causing the supply of bonds to contract and the supply curve to shift to the left. The outcome was then exactly the opposite of that graphed in Figure 5-4 (page 94): the rightward shift of the demand curve and leftward shift of the supply curve led to a rise in the bond price and a fall in interest rates.

The business cycle contraction and the resulting lack of investment opportunities in Japan also led to lower interest rates by decreasing the supply of bonds and shifting the supply curve to the left. Although the demand curve also would shift

to the left because wealth decreased during the business cycle contraction, we have seen in the preceding application that the demand curve would shift less than the supply curve. Thus, the bond price rose and interest rates fell (the opposite outcome to that in Figure 5-6, page 96).

Usually, we think that low interest rates are a good thing because they make it cheap to borrow. But the Japanese example shows that just as there is a fallacy in the adage “You can never be too rich or too thin” (maybe you can’t be too rich, but you can certainly be too thin and do damage to your health), there is a fallacy in always thinking that lower interest rates are better. In Japan, the low and even negative interest rates were a sign that the Japanese economy was in real trouble, with falling prices and a contracting economy. Only when the Japanese economy returns to health will interest rates rise back to more normal levels.

APPLICATION

Have Lower Savings Rates in Canada Led to Higher Interest Rates?

Since 1980, Canada has experienced a sharp drop in personal savings rates, with record lows in recent years. Many commentators, including high officials of the Bank of Canada, have blamed the profligate spending habits of the Canadian public for high interest rates. Are they right?

Our supply and demand analysis of the bond market indicates that they could be right. The decline in savings means that the wealth of Canadian households is lower than would otherwise be the case. This smaller amount of wealth decreases the demand for bonds and shifts the demand curve to the left from B_1^d to B_2^d , as shown in Figure 5-8. The result is that the equilibrium bond price drops from P_1

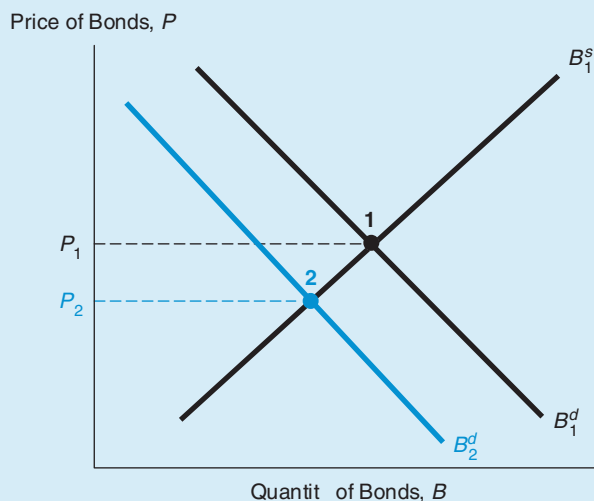


FIG RE 11-8 Response to a Lower Savings Rate

With a lower savings rate, all other things equal, wealth decreases, and the demand curve shifts from B_1^d to B_2^d . The equilibrium moves from point 1 to point 2, with the result that the equilibrium bond price drops from P_1 to P_2 and the equilibrium interest rate rises.

to P_2 and the interest rate rises. Low savings can thus raise interest rates, and the higher rates may retard investment in capital goods. The low savings rate of Canadians may therefore lead to a less-productive economy and is of serious concern to both economists and policymakers. Suggested remedies for the problem range from changing the tax laws to encourage saving to forcing Canadians to save more by mandating increased contributions into retirement plans.

SUPPLY AND DEMAND IN THE MARKET FOR MONEY : THE LIQUIDITY PREFERENCE FRAMEWORK

Instead of determining the equilibrium interest rate using the supply of and demand for bonds, an alternative model developed by John Maynard Keynes, known as the **liquidity preference framework**, determines the equilibrium interest rate in terms of the supply of and demand for money. Although the two frameworks look different, the liquidity preference analysis of the market for money is closely related to the loanable funds framework of the bond market.⁴

The starting point of Keynes's analysis is his assumption that there are two main categories of assets that people use to store their wealth: money and bonds. Therefore, total wealth in the economy must equal the total quantity of bonds plus money in the economy, which equals the quantity of bonds supplied B^s plus the quantity of money supplied M^s . The quantity of bonds B^d and money M^d that people want to hold and thus demand must also equal the total amount of wealth, because people cannot purchase more assets than their available resources allow. The conclusion is that the quantity of bonds and money supplied must equal the quantity of bonds and money demanded:

$$B^s + M^s = B^d + M^d \quad (2)$$

Collecting the bond terms on one side of the equation and the money terms on the other, this equation can be rewritten as

$$B^s - B^d = M^d - M^s \quad (3)$$

The rewritten equation tells us that if the market for money is in equilibrium ($M^s = M^d$), the right-hand side of Equation 3 equals zero, implying that $B^s = B^d$, meaning that the bond market is also in equilibrium.

Thus it is the same to think about determining the equilibrium interest rate by equating the supply and demand for bonds or by equating the supply and demand for money. In this sense, the liquidity preference framework, which analyzes the market for money, is equivalent to a framework analyzing supply and demand in the bond market. In practice, the approaches differ because by assuming that there are only two kinds of assets, money and bonds, the liquidity preference approach implicitly ignores any effects on interest rates that arise from changes in the expected

⁴ Note that the term *market for money* refers to the market for the medium of exchange, money. This market differs from the *money market* referred to by finance practitioners, which is the financial market in which short-term debt instruments are traded.

returns on real assets such as automobiles and houses. In most instances, however, both frameworks yield the same predictions.

The reason that we approach the determination of interest rates with both frameworks is that the bond supply and demand framework is easier to use when analyzing the effects from changes in expected inflation, whereas the liquidity preference framework provides a simpler analysis of the effects from changes in income, the price level, and the supply of money.

Because the definition of money that Keynes used includes currency (which earns no interest) and chequing account deposits (which in his time typically earned little or no interest), he assumed that money has a zero rate of return. Bonds, the only alternative asset to money in Keynes's framework, have an expected return equal to the interest rate i .⁵ As this interest rate rises (holding everything else unchanged), the expected return on money falls relative to the expected return on bonds, and as the theory of asset demand tells us, this causes the demand for money to fall.

We can also see that the demand for money and the interest rate should be negatively related by using the concept of **opportunity cost**, the amount of interest (expected return) sacrificed by not holding the alternative asset—in this case, a bond. As the interest rate on bonds, i , rises, the opportunity cost of holding money rises, and so money is less desirable and the quantity of money demanded must fall.

Figure 5-9 shows the quantity of money demanded at a number of interest rates, with all other economic variables, such as income and the price level, held constant. At an interest rate of 25%, point A shows that the quantity of money

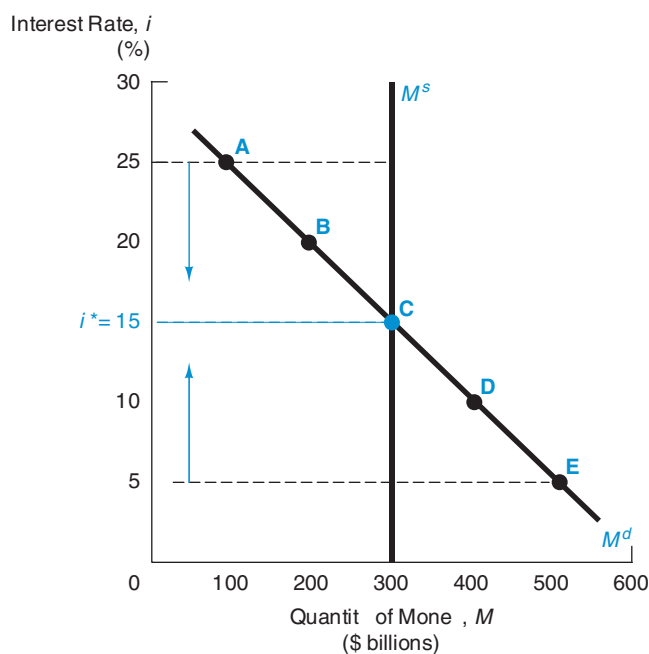


FIG RE 11-9 Equilibrium in the Market for Money

⁵ Keynes did not actually assume that the expected returns on bonds equalled the interest rate but rather argued that they were closely related. This distinction makes no appreciable difference in our analysis.

demand is \$100 billion. If the interest rate is at the lower rate of 20%, the opportunity cost of money is lower, and the quantity of money demanded rises to \$200 billion, as indicated by the move from point A to point B. If the interest rate is even lower, the quantity of money demanded is even higher, as is indicated by points C, D, and E. The curve M^d connecting these points is the demand curve for money, and it slopes downward.

At this point in our analysis, we will assume that a central bank controls the amount of money supplied at a fixed quantity of \$300 billion, so the supply curve for money M^s in the figure is a vertical line at \$300 billion. The equilibrium where the quantity of money demanded equals the quantity of money supplied occurs at the intersection of the supply and demand curves at point C, where

$$M^d = M^s \quad (4)$$

The resulting equilibrium interest rate is at $i^* = 15\%$.

We can again see that there is a tendency to approach this equilibrium by first looking at the relationship of money demand and supply when the interest rate is above the equilibrium interest rate. When the interest rate is 25%, the quantity of money demanded at point A is \$100 billion, yet the quantity of money supplied is \$300 billion. The excess supply of money means that people are holding more money than they desire, so they will try to get rid of their excess money balances by trying to buy bonds. Accordingly, they will bid up the price of bonds, and as the bond price rises, the interest rate will fall toward the equilibrium interest rate of 15%. This tendency is shown by the downward arrow drawn at the interest rate of 25%.

Likewise, if the interest rate is 5%, the quantity of money demanded at point E is \$500 billion, but the quantity of money supplied is only \$300 billion. There is now an excess demand for money because people want to hold more money than they currently have. To try to obtain more money, they will sell their only other asset—bonds—and the price will fall. As the price of bonds falls, the interest rate will rise toward the equilibrium rate of 15%. Only when the interest rate is at its equilibrium value will there be no tendency for it to move further, and the interest rate will settle to its equilibrium value.

CHANGES IN EQUILIBRIUM INTEREST RATES

Analyzing how the equilibrium interest rate changes using the liquidity preference framework requires that we understand what causes the demand and supply curves for money to shift.

Shifts in the Demand for Money

In Keynes's liquidity preference analysis, two factors cause the demand curve for money to shift: income and the price level.

INCOME EFFECT In Keynes's view, there were two reasons why income would affect the demand for money. First, as an economy expands and income rises, wealth increases and people will want to hold more money as a store of value. Second, as the economy expands and income rises, people will want to carry out more transactions using money, with the result that they will also want to hold more money. The conclusion is that ***a higher level of income causes the demand for money at each interest rate to increase and the demand curve to shift to the right.***

PRICE-LEVEL EFFECT Keynes took the view that people care about the amount of money they hold in real terms, that is, in terms of the goods and services that it can buy. When the price level rises, the same nominal quantity of money is no longer as valuable; it cannot be used to purchase as many real goods or services. To restore their holdings of money in real terms to their former level, people will want to hold a greater nominal quantity of money, so **a rise in the price level causes the demand for money at each interest rate to increase and the demand curve to shift to the right.**

Shifts in the Supply of Money

We will assume that the supply of money is completely controlled by the central bank, which in Canada is the Bank of Canada. (Actually, the process that determines the money supply is substantially more complicated, involving banks, depositors, and borrowers from banks. We will study it in more detail later in the book.) For now, all we need to know is that **an increase in the money supply engineered by the Bank of Canada will shift the supply curve for money to the right.**

APPLICATION

Changes in the Equilibrium Interest Rate Due to Changes in Income, the Price Level, or the Money Supply

To see how the liquidity preference framework can be used to analyze the movement of interest rates, we will again look at several applications that will be useful in evaluating the effect of monetary policy on interest rates. In going through these applications, remember to use the *ceteris paribus* assumption: when examining the effect of a change in one variable, hold all other variables constant. (As a study aid, Table 5-4 summarizes the shifts in the demand and supply curves for money.)

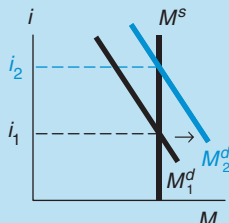
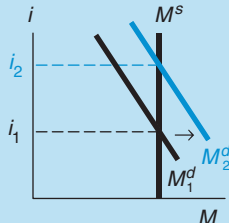
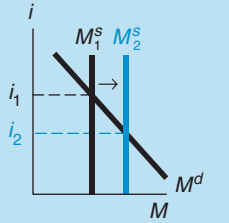
Changes in Income

When income is rising during a business cycle expansion, we have seen that the demand for money will rise, as shown in Figure 5-10 by the shift rightward in the demand curve from M_1^d to M_2^d . The new equilibrium is reached at point 2 at the intersection of the M_2^d curve with the money supply curve M^s . As you can see, the equilibrium interest rate rises from i_1 to i_2 . The liquidity preference framework thus generates the conclusion that **when income is rising during a business cycle expansion (holding other economic variables constant), interest rates will rise.** This conclusion is unambiguous when contrasted to the conclusion reached about the effects of a change in income on interest rates using the bond supply and demand framework.

Changes in the Price Level

When the price level rises, the value of money in terms of what it can purchase is lower. To restore their purchasing power in real terms to its former level, people will want to hold a greater nominal quantity of money. A higher price level shifts the demand curve for money to the right from M_1^d to M_2^d (see Figure 5-10). The equilibrium moves from point 1 to point 2, where the equilibrium interest rate has risen from i_1 to i_2 , illustrating that **when the price level increases, with the supply of money and other economic variables held constant, interest rates will rise.**

TABLE 11-4 Factors That Shift the Demand for and Supply of Money

Variable	Change in Variable	Change in Money Demand (M^d) or Supply (M^s) at Each Interest Rate	Change in Interest Rate	
Income	↑	$M^d \uparrow$	↑	
Price level	↑	$M^d \uparrow$	↑	
Money supply	↑	$M^s \uparrow$	↓	

Note: Only increases (↑) in the variables are shown. The effect of decreases in the variables on the change in demand would be the opposite of those indicated in the remaining columns.

Changes in the Money Supply

An increase in the money supply due to an expansionary monetary policy by the Bank of Canada implies that the supply curve for money shifts to the right. As is shown in Figure 5-11 b the movement of the supply curve from M_1^s to M_2^s , the equilibrium moves from point 1 down to point 2, where the M_2^s supply curve intersects with the demand curve M^d and the equilibrium interest rate has fallen from i_1 to i_2 . **When the money supply increases (everything else remaining equal), interest rates will decline.**⁶

⁶ This same result can be generated using the supply and demand for bonds framework.

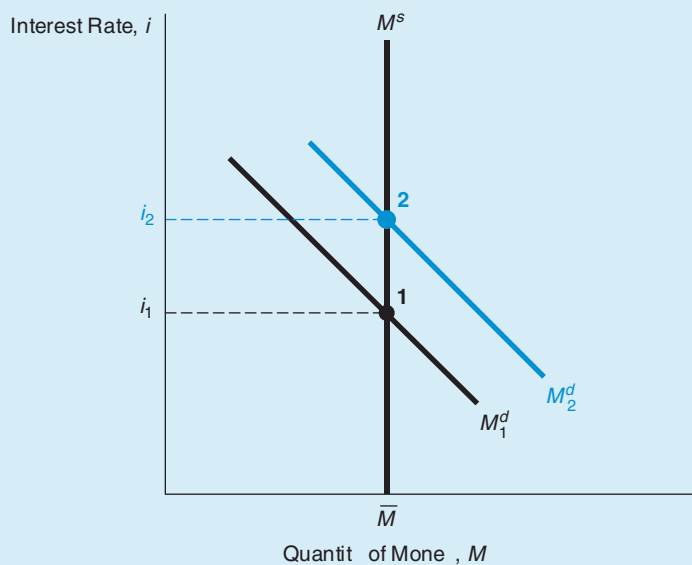


FIG RE 11-10 Response to a Change in Income or the Price Level

In a business cycle expansion, when income is rising, or when the price level rises, the demand curve shifts from M_1^d to M_2^d . The supply curve is fixed at $M^s = \bar{M}$. The equilibrium interest rate rises from i_1 to i_2 .

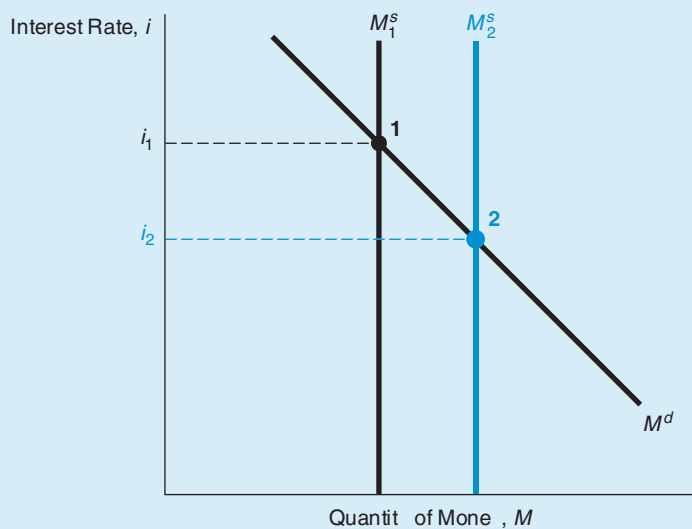


FIG RE 11-11 Response to a Change in the Money Supply

When the money supply increases, the supply curve shifts from M_1^s to M_2^s , and the equilibrium interest rate falls from i_1 to i_2 .

APPLICATION

Money and Interest Rates

The liquidity preference analysis in Figure 5-11 seems to lead to the conclusion that an increase in the money supply will lower interest rates. This conclusion has important policy implications because it has frequently caused politicians to call for a more rapid growth of the money supply in order to drive down interest rates.

But is this conclusion that money and interest rates should be negatively related correct? Might there be other important factors left out of the liquidity preference analysis in Figure 5-11 that would reverse this conclusion? We will provide answers to these questions by applying the supply and demand analysis we have used in this chapter to obtain a deeper understanding of the relationship between money and interest rates.

Milton Friedman, a Nobel laureate in economics, has raised an important criticism of the conclusion that a rise in the money supply lowers interest rates. He acknowledges that the liquidity preference analysis is correct and calls the result—that an increase in the money supply (*everything else remaining equal*) lowers interest rates—the *liquidity effect*. However, he views the liquidity effect as merely part of the story: an increase in the money supply might not leave “everything else equal” and will have other effects on the economy that may make interest rates rise. If these effects are substantial, it is entirely possible that when the money supply rises, interest rates too may rise.

We have already laid the groundwork to discuss these other effects because we have shown how changes in income, the price level, and expected inflation affect the equilibrium interest rate.

1. *Income Effect.* Because an increasing money supply is an expansionary influence on the economy, it should raise national income and wealth. Both the liquidity preference and bond supply and demand frameworks indicate that interest rates will then rise (see Figure 5-6 on page 96 and 5-10 on page 104). Thus ***the income effect of an increase in the money supply is a rise in interest rates in response to the higher level of income.***
2. *Price-Level Effect.* An increase in the money supply can also cause the overall price level in the economy to rise. The liquidity preference framework predicts that this will lead to a rise in interest rates. So ***the price-level effect from an increase in the money supply is a rise in interest rates in response to the rise in the price level.***
3. *Expected-Inflation Effect.* The higher inflation rate that results from an increase in the money supply also affects interest rates by affecting the expected inflation rate. Specifically, an increase in the money supply may lead people to expect a higher price level in the future—hence the expected inflation rate will be higher. The supply and demand for bonds framework has shown us that this increase in expected inflation will lead to a higher level of interest rates. Therefore, ***the expected-inflation effect of an increase in the money supply is a rise in interest rates in response to the rise in the expected inflation rate.***

At first glance it might appear that the price-level effect and the expected-inflation effect are the same thing. They both indicate that increases in the price level induced by an increase in the money supply will raise interest rates. However, there is a subtle difference between the two, and this is why they are discussed as two separate effects.

Suppose that there is a onetime increase in the money supply today that leads to a rise in prices to a permanently higher level by next year. As the price level rises over the course of this year, the interest rate will rise via the price-level effect. Only at the end of the year, when the price level has risen to its peak, will the price-level effect be at a maximum.

The rising price level will also raise interest rates via the expected-inflation effect because people will expect that inflation will be higher over the course of the year. However, when the price level stops rising next year, inflation and the expected inflation rate will return to zero. Any rise in interest rates as a result of the earlier rise in expected inflation will then be reversed. We thus see that in contrast to the price-level effect, which reaches its greatest impact next year, the expected-inflation effect will have its smallest impact (zero impact) next year. The basic difference between the two effects, then, is that the price-level effect remains even after prices have stopped rising, whereas the expected-inflation effect disappears.

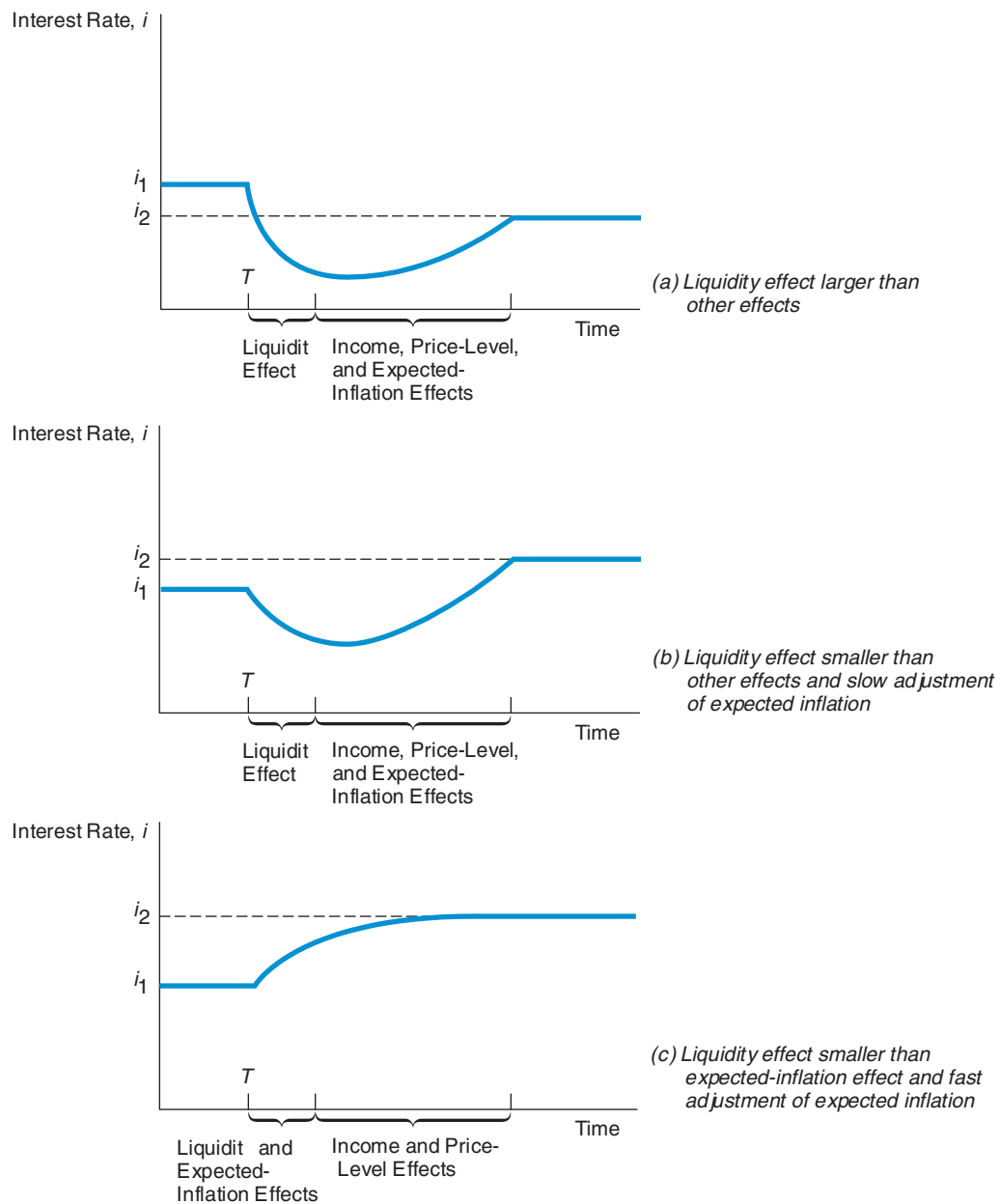
An important point is that the expected-inflation effect will persist only as long as the price level continues to rise. As we will see in our discussion of monetary theory in subsequent chapters, a onetime increase in the money supply will not produce a continually rising price level; only a higher rate of money supply growth will. Thus a higher rate of money supply growth is needed if the expected-inflation effect is to persist.

Does a Higher Rate of Growth of the Money Supply Lower Interest Rates?

We can now put together all the effects we have discussed to help us decide whether our analysis supports the politicians who advocate a greater rate of growth of the money supply when they feel that interest rates are too high. Of all the effects, only the liquidity effect indicates that a higher rate of money growth will cause a decline in interest rates. In contrast, the income, price-level, and expected-inflation effects indicate that interest rates will rise when money growth is higher. Which of these effects are largest, and how quickly do they take effect? The answers are critical in determining whether interest rates will rise or fall when money supply growth is increased.

Generally, the liquidity effect from the greater money growth takes effect immediately because the rising money supply leads to an immediate decline in the equilibrium interest rate. The income and price-level effects take time to work because it takes time for the increasing money supply to raise the price level and income, which in turn raise interest rates. The expected-inflation effect, which also raises interest rates, can be slow or fast, depending on whether people adjust their expectations of inflation slowly or quickly when the money growth rate is increased.

Three possibilities are outlined in Figure 5-12; each shows how interest rates respond over time to an increased rate of money supply growth starting at time T . Panel (a) shows a case in which the liquidity effect dominates the other effects so that the interest rate falls from i_1 at time T to a final level of i_2 . The liquidity effect operates quickly to lower the interest rate, but as time goes by the other effects start to reverse some of the decline. Because the liquidity effect is larger than the others, however, the interest rate never rises back to its initial level.

**FIG RE 11-12** Response over Time to an Increase in Money Supply Growth

Panel (b) has a smaller liquidity effect than the other effects, with the expected-inflation effect operating slowly because expectations of inflation are slow to adjust upward. Initially, the liquidity effect drives down the interest rate. Then the income, price-level, and expected-inflation effects begin to raise it. Because these effects are dominant, the interest rate eventually rises above its initial level i_2 . In the short run, lower interest rates result from increased money growth, but eventually they end up climbing above the initial level.

Panel (c) has the expected-inflation effect dominating as well as operating rapidly because people quickly raise their expectations of inflation when the rate of

money growth increases. The expected-inflation effect begins immediately to overpower the liquidity effect, and the interest rate immediately starts to climb. Over time, as the income and price-level effects start to take hold, the interest rate rises even higher, and the eventual outcome is an interest rate that is substantially above the initial interest rate. The result shows clearly that increasing money supply growth is not the answer to reducing interest rates; rather, money growth should be reduced in order to lower interest rates!

An important issue for economic policymakers is which of these three scenarios is closest to reality. If a decline in interest rates is desired, then an increase in money supply growth is called for when the liquidity effect dominates the other effects, as in panel (a). A decrease in money growth is appropriate if the other effects dominate the liquidity effect and expectations of inflation adjust rapidly, as in panel (c). If the other effects dominate the liquidity effect but expectations of inflation adjust only slowly, as in panel (b), then whether you want to increase or decrease money growth depends on whether you care more about what happens in the short run or the long run.

Which scenario does the evidence support? The relationship of interest rates and money growth from 1968 to 2008 is plotted in Figure 5-13. When the rate of money supply growth began to climb in the late-1970s, interest rates rose, indicating that the price-level, income, and expected-inflation effects dominated the liquidity effect. By the early 1980s, interest rates reached levels unprecedented in the post-World War II period, as did the rate of money supply growth.

The scenario depicted in panel (a) of Figure 5-12 seems doubtful, and the case for lowering interest rates by raising the rate of money growth is much weakened. Looking back at Figure 5-5 (page 95), which shows the relationship between interest rates and expected inflation, you should not find this too surprising. The rise in the rate of money supply growth in the 1960s and 1970s is matched by a large rise in expected inflation, which would lead us to predict that the expected-inflation effect would be dominant. It is the most plausible explanation for why interest rates rose in the face of higher money growth. However, Figure 5-13 does not really tell

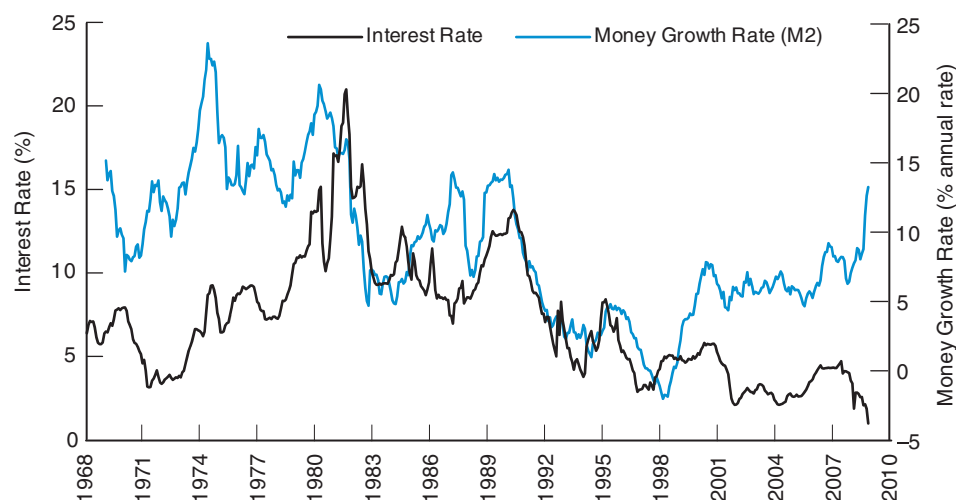


FIG RE 11-13 Money Growth (M2, Annual Rate) and Interest Rates (Three-Month Treasury Bills), 1968–2008

Source: Statistics Canada CANSIM II Series V122531 and V41552796.

us which one of the two scenarios, panel (b) or panel (c) of Figure 5-12, is more accurate. It depends critically on how fast people's expectations about inflation adjust. However, recent research using more sophisticated methods than just looking at a graph like Figure 5-13 do indicate that increased money growth temporarily lowers short-term interest rates (this is also indicated in Figure 5-13 in the 2000s).⁷ However, as you can see in the FYI box Forecasting Interest Rates, interest rate forecasting is a perilous business.

FYI

Forecasting Interest Rates

Forecasting interest rates is a time-honoured profession. Economists are hired (sometimes at very high salaries) to forecast interest rates because businesses need to know what the rates will be in order to plan their future spending, and banks and investors require interest-rate forecasts in order to decide which assets to buy.

The media frequently report interest rate forecasts by leading prognosticators. These forecasts are produced using a wide range of statistical models and a number of different sources of information. One of the most popular methods is based on the bond supply and demand framework described earlier in the chapter. Using this framework, analysts predict what will happen to the factors that affect the supply of and demand for bonds and then use the supply and demand analysis outlined in the chapter to come up with their interest-rate forecasts.

An alternative method of forecasting interest rates makes use of econometric models, models whose equations are estimated with statistical procedures using past data. Many of these econometric models are quite large, involving hundreds and sometimes over a thousand interlocking equations. They produce simultaneous forecasts for many variables, including interest rates, under the assumption that the estimated relationships between variables do not change over time.

Good forecasts of future interest rates are extremely valuable to households and businesses, which, not surprisingly, would be willing to pay a lot for accurate forecasts. Unfortunately, forecasting interest rates is a perilous business. To their embarrassment, even the top experts are frequently far off in their forecasts.

⁷ See Lawrence J. Christiano and Martin Eichenbaum, "Identification and the Liquidity Effect of a Monetary Policy Shock," in *Business Cycles, Growth, and Political Economy*, ed. Alex Cukierman, Zvi Hercowitz, and Leonardo Leiderman (Cambridge, Mass.: MIT Press, 1992), pp. 335–370; Eric M. Leeper and David B. Gordon, "In Search of the Liquidity Effect," *Journal of Monetary Economics* 29 (1992): 341–370; Steven Strongin, "The Identification of Monetary Policy Disturbances: Explaining the Liquidity Puzzle," *Journal of Monetary Economics* 35 (1995): 463–497; Adrian Pagan and John C. Robertson, "Resolving the Liquidity Effect," *Federal Reserve Bank of St. Louis Review* 77 (May–June 1995): 33–54; and Ben S. Bernanke and Illian Mihov, "Measuring Monetary Policy," *Quarterly Journal of Economics* 63 (August 1998): 869–902.

SUMMAR

1. The theory of asset demand tells us that the quantity demanded of an asset is (a) positively related to wealth, (b) positively related to the expected return on the asset relative to alternative assets, (c) negatively related to the riskiness of the asset relative to alternative assets, and (d) positively related to the liquidity of the asset relative to alternative assets.
2. The supply and demand analysis for bonds provides one theory of how interest rates are determined. It predicts that interest rates will change when there is a change in demand because of changes in income (or wealth), expected returns, risk, or liquidity or when there is a change in supply because of changes in the attractiveness of investment opportunities, the real cost of borrowing, or government activities.
3. An alternative theory of how interest rates are determined is provided by the liquidity preference framework, which analyses the supply of and demand for money. It shows that interest rates will change when there is a change in the demand for money because of changes in income or the price level or when there is a change in the supply of money.
4. There are four possible effects of an increase in the money supply on interest rates: the liquidity effect, the income effect, the price-level effect, and the expected-inflation effect. The liquidity effect indicates that a rise in money supply growth will lead to a decline in interest rates; the other effects work in the opposite direction. The evidence seems to indicate that the income, price-level, and expected-inflation effects dominate the liquidity effect such that an increase in money supply growth leads to higher rather than lower interest rates.

KEY TERMS

asset market approach, p. 88	liquidity, p. 83	risk, p. 83
demand curve, p. 85	liquidity preference framework, p. 99	supply curve, p. 86
excess demand, p. 87	market equilibrium, p. 87	theory of asset demand, p. 84
excess supply, p. 87	opportunity cost, p. 100	wealth, p. 83
expected return, p. 83		
Fisher effect, p. 95		

QUESTIONS

You will find the answers to the questions marked with an asterisk in the Textbook Resources section of your MyEconLab.

1. Explain why you would be more or less willing to buy a share of Air Canada stock in the following situations:
 - a. Your wealth falls.
 - b. You expect the stock to appreciate in value.
 - c. The bond market becomes more liquid.
 - d. You expect gold to appreciate in value.
 - e. Prices in the bond market become more volatile.
- *2. Explain why you would be more or less willing to buy a house under the following circumstances:
 - a. You just inherited \$100 000.
 - b. Real estate commissions fall from 6% of the sales price to 5% of the sales price.
 - c. You expect Air Canada stock to double in value next year.
 - d. Prices in the stock market become more volatile.
 - e. You expect housing prices to fall.
3. Explain why you would be more or less willing to buy gold under the following circumstances:
 - a. Gold again becomes acceptable as a medium of exchange.
 - b. Prices in the gold market become more volatile.
 - c. You expect inflation to rise, and gold prices tend to move with the aggregate price level.
 - d. You expect interest rates to rise.
- *4. Explain why you would be more or less willing to buy long-term Air Canada bonds under the following circumstances:
 - a. Trading in these bonds increases, making them easier to sell.
 - b. You expect a bear market in stocks (stock prices are expected to decline).
 - c. Brokerage commissions on stocks fall.
 - d. You expect interest rates to rise.
 - e. Brokerage commissions on bonds fall.
5. What would happen to the demand for Rembrandts if the stock market undergoes a boom? Why?

Answer each question by drawing the appropriate supply and demand diagrams.

- *6. An important way in which the Bank of Canada decreases the money supply is by selling bonds to the public. Using a supply and demand analysis for bonds, show what effect this action has on interest rates. Is your answer consistent with what you would expect to find with the liquidity preference framework?
- 7. Using both the liquidity preference and supply and demand for bonds frameworks, show why interest rates are procyclical (rising when the economy is expanding and falling during recessions).
- *8. Why should a rise in the price level (but not in expected inflation) cause interest rates to rise when the nominal money supply is fixed?
- 9. What effect will a sharp increase in personal savings rates have on Canadian interest rates?
- 10. What effect will a sudden increase in the volatility of gold prices have on interest rates?
- *11. How might a sudden increase in people's expectations of future real estate prices affect interest rates?
- 12. Explain what effect a large federal deficit might have on interest rates.
- *13. Using both the supply and demand for bonds and liquidity preference frameworks, show what the effect is on interest rates when the riskiness of bonds rises. Are the results the same in the two frameworks?

- 14. If the price level falls next year, remaining fixed thereafter, and the money supply is fixed, what is likely to happen to interest rates over the next two years? (Hint: Take account of both the price-level effect and the expected-inflation effect.)
- *15. Will there be an effect on interest rates if brokerage commissions on stocks fall? Explain your answer.

Predicting the Future

- 16. The governor of the Bank of Canada announces in a press conference that he will fight the higher inflation rate with a new anti-inflation program. Predict what will happen to interest rates if the public believes him.
- *17. The governor of the Bank of Canada announces that interest rates will rise sharply next year, and the market believes him. What will happen to today's interest rate on long-term corporate bonds?
- 18. Predict what will happen to interest rates if the public suddenly expects a large increase in stock prices.
- *19. Predict what will happen to interest rates if prices in the bond market become more volatile.
- 20. If the next governor of the Bank of Canada has a reputation for advocating an even slower rate of money growth than the current governor, what will happen to interest rates? Discuss the possible resulting situations.

QUANTITATIVE PROBLEMS

- 1. The demand curve and supply curve for one-year T-bills (with a face value of \$1000) were estimated using the following equations:

$$B^d: \text{Price} = -\frac{2}{5}\text{Quantity} + 940$$

$$B^s: \text{Price} = \text{Quantity} + 100$$

- a. What is the expected equilibrium price and quantity of T-bills in this market?
- b. Given your answer in (a), which is the expected interest rate in this market?
- *2. The demand curve and supply curve for one-year T-bills (with a face value of \$1000) were estimated using the following equations:

$$B^d: \text{Price} = -\frac{2}{5}\text{Quantity} + 940$$

$$B^s: \text{Price} = \text{Quantity} + 100$$

Following a dramatic decline in the value of the stock market, the demand for bonds increased and this

resulted in a parallel shift in the demand curve for bonds, such as the price of bonds at all quantities increased \$100. Assuming no change in the supply function for bonds, what is the new equilibrium price and quantity? What is the new market interest rate?

CANSIM QUESTIONS

- 3. Get the monthly data from 1976 to 2009 on the M2 (gross) monetary aggregate (CANSIM series V41552796) and the three-month T-bill rate (series V122531) from the Textbook Resources area of the MyEconLab.
 - a. Calculate the annual money growth rate, using the formula

$$\mu_t = 100 \times (M_{t+12} - M_t) / M_t$$
 - b. Plot the monetary growth rate, μ_t , and the nominal interest rate, i_t .
 - c. What is the correlation coefficient between the money growth, μ_t , and the nominal interest rate, i_t ? Do you find anything interesting?

4. Get the monthly data from 1976 to 2009 on the three-month T-bill rate (CANSIM series V122531) from the Textbook Resources area of the MyEconLab.
 - a. Plot the nominal interest rate, i .
 - b. Calculate the change in i .

$$\Delta i = i_{+1} - i$$
 - c. Plot Δi . Has the nominal interest rate become more or less volatile over the sample period?

WEB EXERCISES

1. One of the largest single influences on the level of interest rates is inflation. There are a number of sites that report inflation over time. Go to **www.bankofcanada.ca/en/cpi.htm** and review the data. What has the average rate of inflation been since 1995? What year had the lowest level of inflation? What year had the highest level of inflation?
2. Increasing prices erode the purchasing power of the dollar. It is interesting to calculate how much goods would have cost at some point in the past after adjusting for inflation. Click on **www.bankofcanada.ca/en/inflation_calc.htm**. What is the cost today of a car that cost \$10 000 the year that you were born?
3. One of the points made in this chapter is that inflation erodes investment returns. Go to **www.moneychimp.com/articles/econ/inflation_calculator.htm** and review how changes in inflation alter your real return. What happens to the difference between the adjusted value of an investment compared to its inflation-adjusted value as:
 - a. Inflation increases?
 - b. The investment horizon lengthens?
 - c. Expected returns increase?



Be sure to visit the MyEconLab website at **www.myeconlab.com**. This online homework and tutorial system puts you in control of your own learning with study and practice tools directly correlated to this chapter content.

On the MyEconLab website you will find the following appendices and mini-case for this chapter:

Appendix 5.1: Models of Asset Pricing

Appendix 5.2: Applying the Asset Market Approach to a Commodity Market: The Case of Gold

Mini-Case 5.1: The Behaviour of Interest Rates

CHAPTER 12

The Risk and Term Structure of Interest Rates

LEARNING OBJECTIVES

After studying this chapter you should be able to

1. describe how default risk, liquidity, and tax considerations affect interest rates
2. explain how interest rates on bonds with different maturities are related by applying the expectations theory, the segmented markets theory, and the liquidity premium theory
3. predict the movement of short-term interest rates in the future using the yield curve

PREVIEW

In our supply and demand analysis of interest-rate behaviour in Chapter 5, we examined the determination of just one interest rate. Yet we saw earlier that there are enormous numbers of bonds on which the interest rates can and do differ. In this chapter we complete the interest-rate picture by examining the relationship of the various interest rates to one another. Understanding why they differ from bond to bond can help businesses, banks, insurance companies, and private investors decide which bonds to purchase as investments and which ones to sell.

We first look at why bonds with the same term to maturity have different interest rates. The relationship among these interest rates is called the **risk structure of interest rates**, although risk and liquidity both play a role in determining the risk structure. A bond's term to maturity also affects its interest rate, and the relationship among interest rates on bonds with different terms to maturity is called the **term structure of interest rates**. In this chapter we examine the sources and causes of fluctuations in interest rates relative to one another and look at a number of theories that explain these fluctuations.

RISK STRUCTURE OF INTEREST RATES

Figure 6-1 shows the yields to maturity for several categories of long-term bonds from 1978 to 2008. It shows us two important features of interest-rate behaviour for bonds of the same maturity: interest rates on different categories of bonds differ from one another in any given year, and the spread (or difference) between the interest rates varies over time. The interest rates on corporate bonds, for example, are above those on Canada bonds and provincial bonds. In addition, the spread between the interest rates on corporate bonds and Canada bonds is very large during the 1980–1982 and

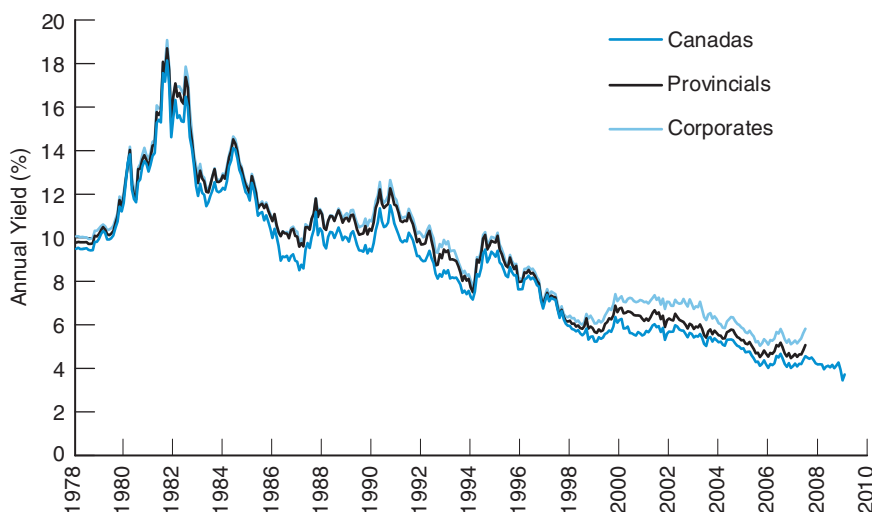


FIG RE 12-1 Long-Term Bond Yields, 1978–2008

Source: Statistics Canada CANSIM II Series V122544, V122517, and V122518.

1990–1991 recessions, is smaller during the mid-1990s, and then widens again afterwards. What factors are responsible for these phenomena?

Default Risk

One attribute of a bond that influences its interest rate is its **risk of default**, which occurs when the issuer of the bond is unable or unwilling to make interest payments when promised or pay off the face value when the bond matures. A corporation suffering big losses might be more likely to suspend interest payments on its bonds. The default risk on its bonds would therefore be quite high. By contrast, Canadian government bonds have usually been considered to have no default risk because the federal government can always increase taxes to pay off its obligations. Bonds like these with no default risk are called **default-free bonds**. The spread between the interest rates on bonds with default risk and default-free bonds, called the **risk premium**, indicates how much additional interest people must earn in order to be willing to hold that risky bond. Our supply and demand analysis of the bond market in Chapter 5 can be used to explain why a bond with default risk always has a positive risk premium and why the higher the default risk is, the larger the risk premium will be.

To examine the effect of default risk on interest rates, let us look at the supply and demand diagrams for the default-free (Canadian government) and corporate long-term bond markets in Figure 6-2. To make the diagrams somewhat easier to read, let's assume that initially corporate bonds have the same default risk as Canada bonds. In this case, these two bonds have the same attributes (identical risk and maturity); their equilibrium prices and interest rates will initially be

equal ($P_1^c = P_1^T$ and $i_1^c = i_1^T$), and the risk premium on corporate bonds ($i_1^c - i_1^T$) will be zero.

If the possibility of a default increases because a corporation begins to suffer large losses, the default risk on corporate bonds will increase, and the expected return on these bonds will decrease. In addition, the corporate bond's return will be more uncertain as well. The theory of asset demand predicts that because the expected return on the corporate bond falls relative to the expected return on the default-free

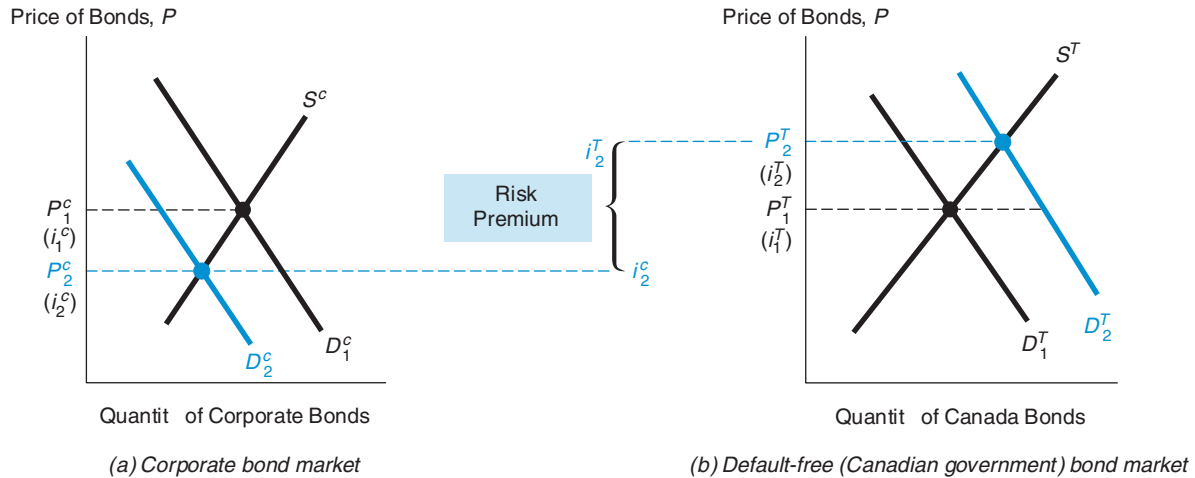


FIG RE 12-2 Response to an Increase in Default Risk on Corporate Bonds

An increase in default risk on corporate bonds shifts the demand curve from D_1^c to D_2^c . Simultaneously, it shifts the demand curve for Canada bonds from D_1^T to D_2^T . The equilibrium price for corporate bonds falls from P_1^c to P_2^c , and the equilibrium interest rate on corporate bonds rises to i_2^c . In the Canada bond market, the equilibrium bond price rises from P_1^T to P_2^T , and the equilibrium interest rate falls to i_2^T . The brace indicates the difference between i_2^c and i_2^T , the risk premium on corporate bonds.

Canada bond while its relative riskiness rises, the corporate bond is less desirable (holding everything else equal), and demand for it will fall. Another way of thinking about this is that if you were an investor, you would want to hold (demand) a smaller amount of corporate bonds. The demand curve for corporate bonds in panel (a) of Figure 6-2 then shifts to the left, from D_1^c to D_2^c .

At the same time, the expected return on default-free Canada bonds increases relative to the expected return on corporate bonds while their relative riskiness declines. The Canada bonds thus become more desirable, and demand rises, as shown in panel (b) by the rightward shift in the demand curve for these bonds from D_1^T to D_2^T .

As we can see in Figure 6-2, the equilibrium price for corporate bonds falls from P_1^c to P_2^c , and since the bond price is negatively related to the interest rate, the equilibrium interest rate on corporate bonds rises to i_2^c . At the same time, however, the equilibrium price for the Canada bonds rises from P_1^T to P_2^T , and the equilibrium interest rate falls to i_2^T . The spread between the interest rates on corporate and default-free bonds—that is, the risk premium on corporate bonds—has risen from zero to $i_2^c - i_2^T$. We can now conclude that **a bond with default risk will always have a positive risk premium, and an increase in its default risk will raise the risk premium.**

Because default risk is so important to the size of the risk premium, purchasers of bonds need to know whether a corporation is likely to default on its bonds. This information is provided by **credit-rating agencies**, investment advisory firms that rate the quality of corporate and municipal bonds in terms of the probability of default. Table 6-1 provides the ratings and their description for the two largest credit-rating agencies, Dominion Bond Rating Service (DBRS) and Standard & Poor's Corporation (S&P)—in the United States, Moody's Investor Service and Standard & Poor's Corporation provide similar information. Bonds with relatively low risk of default are called *investment-grade* securities and have a rating of BBB

TABLE 12-1 Bond Ratings by Standard & Poor's and DBRS

Rating	Definitions
AAA	Highest quality
AA	Superior quality
A	Satisfactory quality
BBB	Adequate quality
BB	Speculative
B	Highly speculative
CCC, CC, C	Very highly speculative
D	In default

and above. Bonds with ratings below BBB have higher default risk and have been aptly dubbed speculative-grade or **junk bonds**. Because these bonds always have higher interest rates than investment-grade securities, they are also referred to as high-yield bonds. Investment-grade securities whose rating has fallen to junk levels are referred to as **fallen angels**.

Next let's look back at Figure 6-1 and see if we can explain the relationship between interest rates on corporate and Canada bonds. Corporate bonds always have higher interest rates than Canada bonds because they always have some risk of default, whereas Canada bonds do not. Because corporate bonds have a greater default risk than Canada bonds, their risk premium is greater, and the corporate bond rate therefore always exceeds the Canada bond rate. We can use the same analysis to explain the huge jump in the risk premium on corporate bond rates during the 1980–1982, 1990–1991, and 2000 recessions (Figure 6-3). The recession periods saw a very high rate of business failures and defaults. As we would expect, these factors led to a substantial increase in default risk for bonds issued by vulnerable corporations, and the risk premium for corporate bonds reached unprecedented high levels.

**FIGURE 12-3** Corporates Canadas Spread_1978

2008 Source: Statistics Canada CANSIM II Series V122518 and V122544.

APPLICATION

The Subprime Collapse and the BAA-Treasury Spread in the United States

Starting in August 2007, the collapse of the subprime mortgage market in the U.S. led to large losses in American financial institutions (which will be discussed more extensively in Chapter 9). As a consequence of the subprime collapse, many investors began to doubt the financial health of corporations with low credit ratings, such as BAA, and even the reliability of the ratings themselves. The perceived increase in default risk for BAA bonds made them less desirable at any given interest rate, decreased the quantity demanded, and shifted the demand curve for BAA bonds to the left. As shown in panel (a) of Figure 6-2, the interest rate on BAA bonds should have risen, which is indeed what happened. Interest rates on BAA bonds rose by 280 **basis points** (2.80 percentage points) from 6.63% at the end of July 2007 to 9.43% at the most virulent stage of the crisis in mid October 2008. But the increase in perceived default risk for BAA bonds after the subprime collapse made default-free U.S. Treasury bonds relatively more attractive and shifted the demand curve for these securities to the right—an outcome described by some analysts as a “flight to quality.” Just as our analysis predicts in Figure 6-2, interest rates on U.S. Treasury bonds fell by 80 basis points, from 4.78% at the end of July 2007 to 3.98% in mid-October 2008. The spread between interest rates on BAA and Treasury bonds rose by 360 basis points from 1.85% before the crisis to 5.45% afterward.

Liquidity

Another attribute of a bond that influences its interest rate is its liquidity. As we learned in Chapter 5, a liquid asset is one that can be quickly and cheaply converted into cash if the need arises. The more liquid an asset is, the more desirable it is (holding everything else constant). Canada bonds are the most liquid of all long-term bonds because they are so widely traded that they are the easiest to sell quickly and the cost of selling them is low. Corporate bonds are not as liquid because fewer bonds for any one corporation are traded; thus it can be costly to sell these bonds in an emergency because it may be hard to find buyers quickly.

How does the reduced liquidity of corporate bonds affect their interest rates relative to the interest rate on Canada bonds? We can use supply and demand analysis with the same figure that was used to analyze the effect of default risk, Figure 6-2, to show that the lower liquidity of corporate bonds relative to Canada bonds increases the spread between the interest rates on these two bonds. Let us start the analysis by assuming that initially corporate and Canada bonds are equally liquid and all their other attributes are the same. As shown in Figure 6-2, their equilibrium prices and interest rates will initially be equal: $P_1^c = P_1^T$ and $i_1^c = i_1^T$. If the corporate bond becomes less liquid than the Canada bond because it is less widely traded, then as the theory of asset demand indicates, its demand will fall, shifting its demand curve from D_1^c to D_2^c , as in panel (a). The Canada bond now becomes relatively more liquid in comparison with the corporate bond, so its demand curve shifts rightward from D_1^T to D_2^T , as in panel (b). The shifts in the curves in Figure 6-2 show that the price of the less-liquid corporate bond falls and its interest rate rises, while the price of the more-liquid Canada bond rises and its interest rate falls.

The result is that the spread between the interest rates on the two bond types has risen. Therefore, the differences between interest rates on corporate bonds and Canada bonds (that is, the risk premiums) reflect not only the corporate bond's default risk but its liquidity too. This is why a risk premium is more accurately a "risk and liquidity premium," but convention dictates that it be called a *risk premium*.

Income Tax Considerations

In Canada, coupon payments on fixed-income securities are taxed as ordinary income in the year they are received. In some other countries, however, certain government bonds are not taxable. In the United States, for example, interest payments on municipal bonds are exempt from federal income taxes, and these bonds have had lower interest rates than U.S. Treasury bonds for at least 40 years. How does taxation affect the interest rate on bonds?

Let us imagine that you have a high enough income to put you in the 40% income tax bracket, where for every extra dollar of income you have to pay 40 cents to the government. If you own a \$1000-face-value taxable bond that sells for \$1000 and has a coupon payment of \$100, you get to keep only \$60 of the payment after taxes. Although the bond has a 10% interest rate, you actually earn only 6% after taxes.

Suppose, however, that you put your savings into a \$1000-face-value tax-exempt bond that sells for \$1000 and pays only \$80 in coupon payments. Its interest rate is only 8%, but because it is a tax-exempt security, you pay no taxes on the \$80 coupon payment, so you earn 8% after taxes. Clearly, you earn more on the tax-exempt bond, so you are willing to hold the bond even though it has a lower interest rate than the taxable bond. Notice that the tax-exempt status of a bond becomes a significant advantage when income tax rates are very high.

APPLICATION

Tax-Exempt versus Taxable Bonds

Suppose you had the opportunity to buy either a tax-exempt bond or a taxable bond, both of which have a face value and purchase price of \$1000. Assume both bonds have identical risk. The tax-exempt bond has coupon payments of \$60 and a coupon rate of 6%. The taxable bond has coupon payments of \$80 and an interest rate of 8%. Which bond would you choose to purchase, assuming a 40% tax rate?

Solution

You would choose to purchase the tax-exempt bond because it will earn you \$60 in coupon payments and an interest rate after taxes of 6%. In this case, you pay no taxes on the \$60 coupon payments and earn 6% after taxes. However, you have to pay taxes on taxable bonds. You will keep only 60% of the \$80 coupon payment because the other 40% goes to taxes. Therefore, you receive \$48 of the coupon payment and have an interest rate of 4.8% after taxes. Buying the tax-exempt bond would yield you higher earnings.

APPLICATION

Effects of the Bush Tax Cut on Bond Interest Rates in the United States

The Bush tax cut passed in 2001 scheduled a reduction of the top income tax bracket from 39% to 35% over a ten-year period. What is the effect of this income tax decrease on interest rates in the municipal bond market relative to those in the U.S. Treasury bond market?

A supply and demand analysis similar to that in Figure 6-2 provides the answer. A decreased income tax rate for wealthy people means that the after-tax expected return on tax-free municipal bonds relative to that on U.S. Treasury bonds is lower because the interest on Treasury bonds is now taxed at a lower rate. Because municipal bonds now become less desirable, their demand decreases, shifting the demand curve to the left, which lowers their price and raises their interest rate. Conversely, the lower income tax rate makes U.S. Treasury bonds more desirable; this change shifts their demand curve to the right, raises their price, and lowers their interest rates.

Our analysis thus shows that the Bush tax cut has caused interest rates on municipal bonds to rise relative to interest rates on Treasury bonds. With the possible repeal of the Bush tax cuts that may occur with the Obama administration, this analysis would be reversed. Higher tax rates would raise the after-tax expected return on tax-free municipal bonds relative to Treasuries. Demand for municipal bonds would increase, shifting the demand curve to the right, which would raise their price and lower their interest rate. Conversely, the higher tax rate would make Treasury bonds less desirable, shifting their demand curve to the left, lowering their price, and raising their interest rate. Higher tax rates would thus result in lower interest rates on municipal bonds relative to the interest rate on Treasury bonds.

Summary

In general, the risk structure of interest rates (the relationship among interest rates on bonds with the same maturity) is explained by three factors: default risk, liquidity, and the income tax treatment of the bond's interest payments. As a bond's default risk increases, the risk premium on that bond (the spread between its interest rate and the interest rate on a default-free Canadian government bond) rises. The greater liquidity of Canada bonds also explains why their interest rates are lower than interest rates on less liquid bonds. If a bond has a favourable tax treatment, as do municipal bonds in the United States whose interest payments are exempt from federal income taxes, its interest rate will be lower.

TERM STRUCTURE OF INTEREST RATES

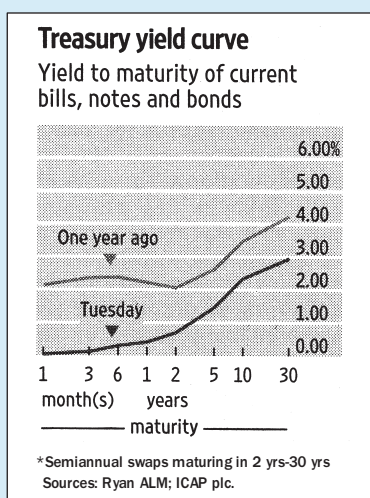
We have seen how risk, liquidity, and tax considerations (collectively embedded in the risk structure) can influence interest rates. Another factor that influences the interest rate on a bond is its term to maturity: bonds with identical risk, liquidity, and tax characteristics may have different interest rates because the time remaining to maturity is different. A plot of the yields on bonds with differing terms to maturity but the same risk, liquidity, and tax considerations is called a **yield curve**, and it describes the term structure of interest rates for particular types of bonds, such as government bonds. The Financial News box Yield Curves shows several

FINANCIAL NEWS

Yield Cur es

The *Wall Street Journal* publishes a daily plot of the yield curves for U.S. Treasury securities, an example of which is presented here. It is found in the Money and Investing section. The *Globe and Mail: Report on Business* publishes similar yield curves for Government of Canada securities.

The numbers on the vertical axis indicate the interest rate for the U.S. Treasury security, with the maturity given by the numbers on the horizontal axis.



Source: *Wall Street Journal*, Wednesday, January 21, 2009, p. C4.

yield curves for U.S. Treasury securities that were published in the *Wall Street Journal*. Similar yield curves are reported for Canada in the *Globe and Mail: Report on Business*. Yield curves can be classified as upward-sloping, flat, and downward-sloping (the last sort is often referred to as an **inverted yield curve**). When yield curves slope upward, as in the Financial News box, the long-term interest rates are above the short-term interest rates; when yield curves are flat, short- and long-term interest rates are the same; and when yield curves are inverted, long-term interest rates are below short-term interest rates. Yield curves can also have more complicated shapes in which they first slope up and then down, or vice versa. Why do we usually see upward slopes of the yield curve?

Besides explaining why yield curves take on different shapes at different times, a good theory of the term structure of interest rates must explain the following three important empirical facts:

1. As we see in Figure 6-4, interest rates on bonds of different maturities move together over time.
2. When short-term interest rates are low, yield curves are more likely to have an

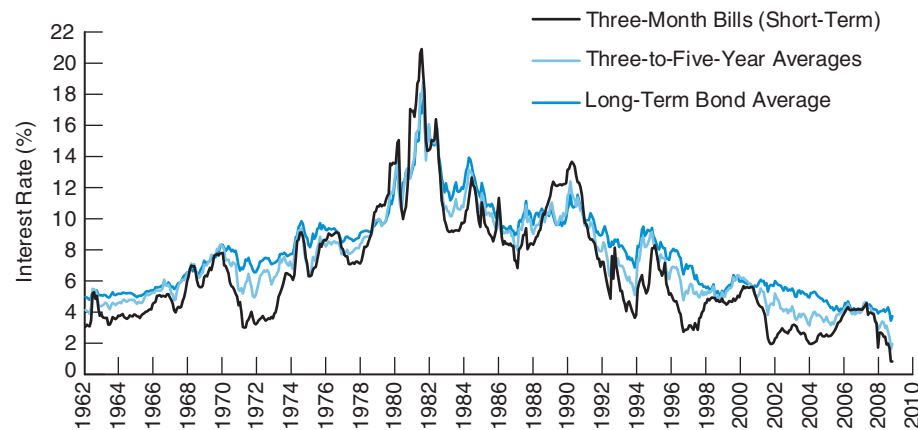


FIG RE 6-4 Movements over Time of Interest Rates on Government of Canada Bonds with Different Maturities, 1962–2008

Source: Statistics Canada CANSIM II Series V122531, V122485, and V122487.

upward slope; when short-term interest rates are high, yield curves are more likely to slope downward and be inverted.

3. Yield curves almost always slope upward, like the yield curves in the Financial News box.

Four theories have been put forward to explain the term structure of interest rates, that is, the relationship among interest rates on bonds of different maturities reflected in yield-curve patterns: (1) the expectations theory, (2) the segmented markets theory, (3) the liquidity premium theory, and (4) the preferred habitat theory. The expectations theory does a good job of explaining the first two facts on our list but not the third. The segmented markets theory can explain fact 3 but not the other two facts, which are well explained by the expectations theory. Because each theory explains facts that the others cannot, a natural way to seek a better understanding of the term structure is to combine features of all four theories, which leads us to the liquidity premium and preferred habitat theories, which can explain all three facts.

If the liquidity premium and preferred habitat theories do a better job of explaining the facts and are hence the most widely accepted theories, why do we spend time discussing the other two theories? There are two reasons. First, the ideas in these two theories provide the groundwork for the liquidity premium and preferred habitat theories. Second, it is important to see how economists modify theories to improve them when they find that the predicted results are inconsistent with the empirical evidence.

Expectations Theory

The **expectations theory** of the term structure states the following common-sense proposition: the interest rate on a long-term bond will equal an average of short-term interest rates that people expect to occur over the life of the long-term bond. For example, if people expect that short-term interest rates will be 10% on average over the coming five years, the expectations theory predicts that the interest rate on bonds with five years to maturity will be 10% too. If short-term interest rates were expected to rise even higher after this five-year

period so that the average short-term interest rate over the coming 20 years is 11%, then the interest rate on 20-year bonds would equal 11% and would be higher than the interest rate on five-year bonds. We can see that the explanation provided by the expectations theory for why interest rates on bonds of different maturities differ is that short-term interest rates are expected to have different values at future dates.

The key assumption behind this theory is that buyers of bonds do not prefer bonds of one maturity over another, so the theory will not hold if the quantity of a bond if its expected return is less than that of another bond with a different maturity. Bonds that have this characteristic are said to be *perfect substitutes*. What this means in practice is that if bonds with different maturities are perfect substitutes, the expected return on these bonds must be equal.

To see how the assumption that bonds with different maturities are perfect substitutes leads to the expectations theory, let us consider the following two investment strategies:

1. Purchase a one-year bond, and when it matures in one year, purchase another one-year bond.
2. Purchase a two-year bond and hold it until maturity.

Because both strategies must have the same expected return if people are holding both one- and two-year bonds, the interest rate on the two-year bond must equal the average of the two one-year interest rates.

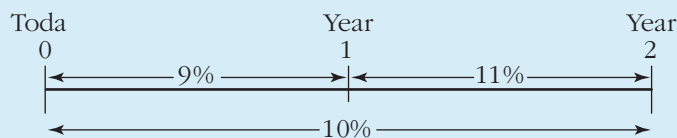
APPLICATION

Expectations Theory

The current interest rate on a one-year bond is 9%, and you expect the interest rate on the one-year bond next year to be 11%. What is the expected return over the two years? What interest rate must a two-year bond have to equal the two one-year bonds?

Solution

The expected return over the two years will average 10% per year ($[9\% + 11\%]/2 = 10\%$). The bondholder will be willing to hold both the one- and two-year bonds only if the expected return per year of the two-year bond equals 10%. Therefore, the interest rate on the two-year bond must equal 10%, the average interest rate on the two one-year bonds. Graphically, we have:



We can make this argument more general. For an investment of \$1, consider the choice of holding, for two periods, a two-period bond or two one-period bonds. Using the definitions

i_t = today's (time t) interest rate on a one-period bond

i_{t+1}^e = interest rate on a one-period bond expected for next period (time $t + 1$)

i_{2t} = today's (time t) interest rate on the two-period bond

the expected return over the two periods from investing \$1 in the two-period bond and holding it for the two periods can be calculated as

$$(1 + i_{2t})(1 + i_{2t}) - 1 = 1 + 2i_{2t} + (i_{2t})^2 - 1 = 2i_{2t} + (i_{2t})^2$$

After the second period, the \$1 investment is worth $(1 + i_{2t})(1 + i_{2t})$. Subtracting the \$1 initial investment from this amount and dividing by the initial \$1 investment gives the rate of return calculated in the above equation. Because $(i_{2t})^2$ is extremely small—if $i_{2t} = 10\% = 0.10$, then $(i_{2t})^2 = 0.01$ —we can simplify the expected return for holding the two-period bond for the two periods to

$$2i_{2t}$$

With the other strategy, in which one-period bonds are bought, the expected return on the \$1 investment over the two periods is

$$(1 + i_t)(1 + i_{t+1}^e) - 1 = 1 + i_t + i_{t+1}^e + i_t(i_{t+1}^e) - 1 = i_t + i_{t+1}^e + i_t(i_{t+1}^e)$$

This calculation is derived by recognizing that after the first period, the \$1 investment becomes $1 + i_t$, and this is reinvested in the one-period bond for the next period, yielding an amount $(1 + i_t)(1 + i_{t+1}^e)$. Then, subtracting the \$1 initial investment from this amount and dividing by the initial investment of \$1 gives the expected return for the strategy of holding one-period bonds for the two periods. Because $i_t(i_{t+1}^e)$ is also extremely small—if $i_t = i_{t+1}^e = 0.10$, then $i_t(i_{t+1}^e) = 0.01$ —we can simplify this to

$$i_t + i_{t+1}^e$$

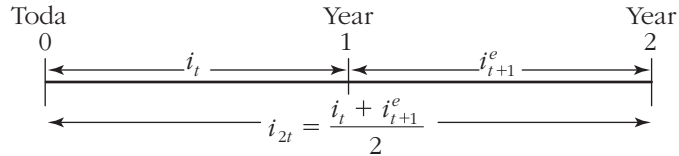
Both bonds will be held only if these expected returns are equal, that is, when

$$2i_{2t} = i_t + i_{t+1}^e$$

Solving for i_{2t} in terms of the one-period rates, we have

$$i_{2t} = \frac{i_t + i_{t+1}^e}{2} \quad (1)$$

which tells us that the two-period rate must equal the average of the two one-period rates. Graphically, this can be shown as:



We can conduct the same steps for bonds with a longer maturity so that we can examine the whole term structure of interest rates. Doing so, we will find that the interest rate of i_n on an n -period bond must equal

$$i_n = \frac{i + i^e_{+1} + i^e_{+2} + \cdots + i^e_{+(n-1)}}{n} \quad (2)$$

Equation 2 states that the n -period interest rate equals the average of the one-period interest rates expected to occur over the n -period life of the bond. This is a restatement of the expectations theorem in more precise terms.¹

APPLICATION

Expectations Theorem and the Yield Curve

The one-year interest rate over the next five years is expected to be 5%, 6%, 7%, 8%, and 9%. Given this information, what are the interest rates on a two-year bond and a five-year bond? Explain what is happening to the yield curve.

Solution

The interest rate on the two-year bond would be 5.5%.

$$i_n = \frac{i + i^e_{+1} + i^e_{+2} + \cdots + i^e_{+(n-1)}}{n}$$

where

$$i = \text{year 1 interest rate} = 5\%$$

$$i^e_{+1} = \text{year 2 interest rate} = 6\%$$

$$n = \text{number of years} = 2$$

Thus

$$i_2 = \frac{5\% + 6\%}{2} = 5.5\%$$

The interest rate on the five-year bond would be 7%.

$$i_n = \frac{i + i^e_{+1} + i^e_{+2} + \cdots + i^e_{+(n-1)}}{n}$$

where

$$i = \text{year 1 interest rate} = 5\%$$

$$i^e_{+1} = \text{year 2 interest rate} = 6\%$$

$$i^e_{+2} = \text{year 3 interest rate} = 7\%$$

$$i^e_{+3} = \text{year 4 interest rate} = 8\%$$

$$i^e_{+4} = \text{year 5 interest rate} = 9\%$$

$$n = \text{number of years} = 5$$

¹ The analysis here has been conducted for discount bonds. Formulas for interest rates on coupon bonds would differ slightly from those used here but would convey the same principle.

Thus

$$i_{5t} = \frac{5\% + 6\% + 7\% + 8\% + 9\%}{5} = 7.0\%$$

Using the same equation for the one-, three-, and four-year interest rates, you will be able to verify the one-year to five-year rates as 5.0%, 5.5%, 6.0%, 6.5%, and 7.0%, respectively. The rising trend in short-term interest rates produces an upward-sloping yield curve along which interest rates rise as maturity lengthens.

The expectations theory is an elegant theory that explains why the term structure of interest rates (as represented by yield curves) changes at different times. When the yield curve is upward-sloping, the expectations theory suggests that short-term interest rates are expected to rise in the future, as we have seen in our numerical example. In this situation, in which the long-term rate is currently above the short-term rate, the average of future short-term rates is expected to be higher than the current short-term rate, which can occur only if short-term interest rates are expected to rise. This is what we see in our numerical example. When the yield curve is inverted (slopes downward), the average of future short-term interest rates is expected to be below the current short-term rate, implying that short-term interest rates are expected to fall, on average, in the future. Only when the yield curve is flat does the expectations theory suggest that short-term interest rates are not expected to change, on average, in the future.

The expectations theory also explains fact 1, that interest rates on bonds with different maturities move together over time. Historically, short-term interest rates have had the characteristic that if they increase today, they will tend to be higher in the future. Hence, a rise in short-term rates will raise people's expectations of future short-term rates. Because long-term rates are the average of expected future short-term rates, a rise in short-term rates will also raise long-term rates, causing short- and long-term rates to move together.

The expectations theory also explains fact 2, that yield curves tend to have an upward slope when short-term interest rates are low and are inverted when short-term rates are high. When short-term rates are low, people generally expect them to rise to some normal level in the future, and the average of future expected short-term rates is high relative to the current short-term rate. Therefore, long-term interest rates will be substantially above current short-term rates, and the yield curve would then have an upward slope. Conversely, if short-term rates are high, people usually expect them to come back down. Long-term rates would then drop below short-term rates because the average of expected future short-term rates would be below current short-term rates and the yield curve would slope downward and become inverted.²

² The expectations theory explains another important fact about the relationship between short-term and long-term interest rates. As you can see looking back at Figure 6-4, short-term interest rates are more volatile than long-term rates. If interest rates are *mean-reverting*—that is, if they tend to head back down after they are at unusually high levels or go back up when they are at unusually low levels—then an average of these short-term rates must necessarily have lower volatility than the short-term rates themselves. Because the expectations theory suggests that the long-term rate will be an average of future short-term rates, it implies that the long-term rate will have lower volatility than short-term rates.

The expectations theory is an attractive theory because it provides a simple explanation of the behaviour of the term structure, but unfortunately it has a major shortcoming: it cannot explain fact 3, that yield curves usually slope upward. The typical upward slope of yield curves implies that short-term interest rates are usually expected to rise in the future. In practice, short-term interest rates are just as likely to fall as they are to rise, and so the expectations theory suggests that the typical yield curve should be flat rather than upward-sloping.

Segmented Markets Theory

As the name suggests, the **segmented markets theory** of the term structure sees markets for different-maturity bonds as completely separate and segmented. The interest rate for each bond with a different maturity is then determined by the supply of and demand for that bond with no effects from expected returns on other bonds with other maturities.

The key assumption in the segmented markets theory is that bonds of different maturities are not substitutes at all, so the expected return from holding a bond of one maturity has no effect on the demand for a bond of another maturity. This theory of the term structure is at the opposite extreme to the expectations theory, which assumes that bonds of different maturities are perfect substitutes.

The argument for why bonds of different maturities are not substitutes is that investors have very strong preferences for bonds of one maturity but not for another, so they will be concerned with the expected returns only for bonds of the maturity they prefer. This might occur because they have a particular holding period in mind, and if they match the maturity of the bond to the desired holding period, they can obtain a certain return with no risk at all.³ (We have seen in Chapter 4 that if the term to maturity equals the holding period, the return is known for certain because it equals the yield exactly, and there is no interest-rate risk.) For example, people who have a short holding period would prefer to hold short-term bonds. Conversely, if you were putting funds away for your young child to go to college, your desired holding period might be much longer, and you would want to hold longer-term bonds.

In the segmented markets theory, differing yield curve patterns are accounted for by supply and demand differences associated with bonds of different maturities. If, as seems sensible, investors have short desired holding periods and generally prefer bonds with shorter maturities that have less interest-rate risk, the segmented markets theory can explain fact 3, that yield curves typically slope upward. Because in the typical situation the demand for long-term bonds is relatively lower than that for short-term bonds, long-term bonds will have lower prices and higher interest rates, and hence the yield curve will typically slope upward.

Although the segmented markets theory can explain why yield curves usually tend to slope upward, it has a major flaw in that it cannot explain facts 1 and 2. Because it views the market for bonds of different maturities as completely segmented, there is no reason for a rise in interest rates on a bond of one maturity to

³ The statement that there is no uncertainty about the return if the term to maturity equals the holding period is literally true only for a discount bond. For a coupon bond with a long holding period, there is some risk because coupon payments must be reinvested before the bond matures. Our analysis here is thus being conducted for discount bonds. However, the gist of the analysis remains the same for coupon bonds because the amount of this risk from reinvestment is small when coupon bonds have the same term to maturity as the holding period.

affect the interest rate on a bond of another maturity. Therefore, it cannot explain why interest rates on bonds of different maturities tend to move together (fact 1). Second, because it is not clear how demand and supply for short- versus long-term bonds change with the level of short-term interest rates, the theory cannot explain why yield curves tend to slope upward when short-term interest rates are low and to be inverted when short-term interest rates are high (fact 2).

Because each of our two theories explains empirical facts that the other cannot, a logical step is to combine the theories, which leads us to the liquidity premium and preferred habitat theories.

Liquidity Premium and Preferred Habitat Theories

The **liquidity premium theory** of the term structure states that the interest rate on a long-term bond will equal an average of short-term interest rates expected to occur over the life of the long-term bond plus a liquidity premium (also referred to as a term premium) that responds to supply and demand conditions for that bond.

The liquidity premium theory's key assumption is that bonds of different maturities are substitutes, which means that the expected return on one bond *does* influence the expected return on a bond of a different maturity, but it allows investors to prefer one bond maturity over another. In other words, bonds of different maturities are assumed to be substitutes but not perfect substitutes. Investors tend to prefer shorter-term bonds because these bonds bear less interest-rate risk. For these reasons, investors must be offered a positive liquidity premium to induce them to hold longer-term bonds. Such an outcome would modify the expectations theory by adding a positive liquidity premium to the equation that describes the relationship between long- and short-term interest rates. The liquidity premium theory is thus written as:

$$i_{nt} = \frac{i_t + i_{t+1}^e + i_{t+2}^e + \cdots + i_{t+(n-1)}^e}{n} + l_{nt} \quad (3)$$

where l_{nt} = the liquidity (term) premium for the n -period bond at time t , which is always positive and rises with the term to maturity of the bond, n .

Closely related to the liquidity premium theory is the **preferred habitat theory**, which takes a somewhat less-direct approach to modifying the expectations hypothesis but comes up with a similar conclusion. It assumes that investors have a preference for bonds of one maturity over another, a particular bond maturity (preferred habitat) in which they prefer to invest. Because they prefer bonds of one maturity over another, they will be willing to buy bonds that do not have the preferred maturity (habitat) only if they earn a somewhat higher expected return. Because investors are likely to prefer the habitat of short-term bonds to that of longer-term bonds, they are willing to hold long-term bonds only if they have higher expected returns. This reasoning leads to the same Equation 3 implied by the liquidity premium theory with a term premium that typically rises with maturity.

The relationship between the expectations theory and the liquidity premium and preferred habitat theories is shown in Figure 6-5. There we see that because the liquidity premium is always positive and typically grows as the term to maturity increases, the yield curve implied by the liquidity premium and preferred habitat theories is always above the yield curve implied by the expectations theory and has a steeper slope. (Note that for simplicity we are assuming that the expectations theory yield curve is flat.)

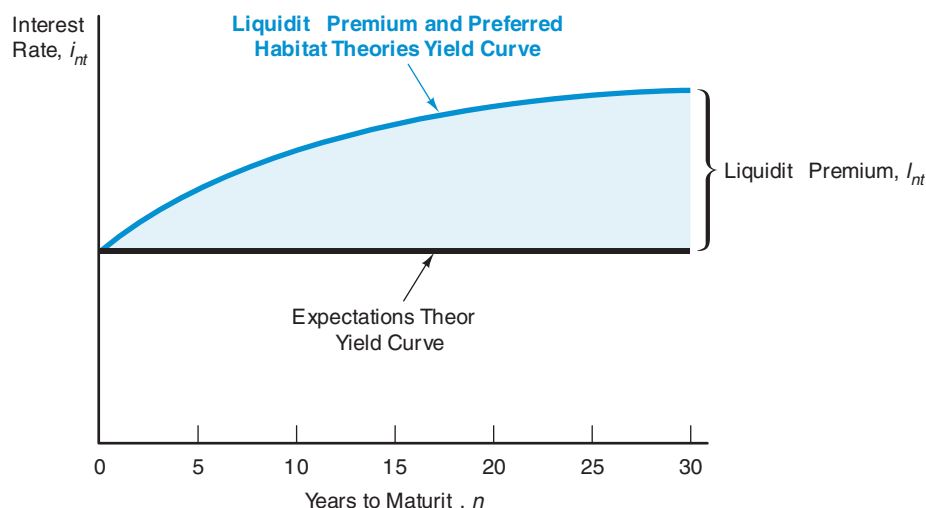


FIG RE 12-5 The Relationship Between the Liquidity Premium and Preferred Habitat Theories and Expectations Theory

Because the liquidity premium is always positive and grows as the term to maturity increases, the yield curve implied by the liquidity premium and preferred habitat theories is always above the yield curve implied by the expectations theory and has a steeper slope. Note that the yield curve implied by the expectations theory is drawn under the scenario of unchanging future one-year interest rates.

APPLICATION

Liquidity Premium Theory

Let's suppose that the one-year interest rate over the next five years is expected to be 5%, 6%, 7%, 8%, and 9%. Investors' preferences for holding short-term bonds have the liquidity premiums for one-year to five-year bonds as 0%, 0.25%, 0.5%, 0.75%, and 1.0%, respectively. What is the interest rate on a two-year bond and a five-year bond? Compare these findings with the answer in the previous Application dealing with the pure expectations theory.

Solution

The interest rate on the two-year bond would be 5.75%.

$$i_n = \frac{i + i^e_{+1} + i^e_{+2} + \cdots + i^e_{+(n-1)}}{n} + l_n$$

where

i = year 1 interest rate = 5%

i^e_{+1} = year 2 interest rate = 6%

l_2 = liquidity premium = 0.25%

n = number of years = 2

Thus

$$i_2 = \frac{5\% + 6\%}{2} + 0.25\% = 5.75\%$$

The interest rate on the five-year bond would be 8%.

$$i_{nt} = \frac{i_t + i_{t+1}^e + i_{t+2}^e + \cdots + i_{t+(n-1)}^e}{n} + l_{nt}$$

where

$$i_t = \text{year 1 interest rate} = 5\%$$

$$i_{t+1}^e = \text{year 2 interest rate} = 6\%$$

$$i_{t+2}^e = \text{year 3 interest rate} = 7\%$$

$$i_{t+3}^e = \text{year 4 interest rate} = 8\%$$

$$i_{t+4}^e = \text{year 5 interest rate} = 9\%$$

$$l_{5t} = \text{liquidity premium} = 1\%$$

$$n = \text{number of years} = 5$$

Thus

$$i_{5t} = \frac{5\% + 6\% + 7\% + 8\% + 9\%}{5} + 1\% = 8.0\%$$

If you did similar calculations for the one-, three-, and four-year interest rates, the one-year to five-year interest rates would be as follows: 5.0%, 5.75%, 6.5%, 7.25%, and 8.0%, respectively. Comparing these findings with those for the pure expectations theory, we can see that the liquidity preference theory produces yield curves that slope more steeply upward because of investors' preferences for short-term bonds.

Let's see if the liquidity premium and preferred habitat theories are consistent with all three empirical facts we have discussed. They explain fact 1, that interest rates on different-maturity bonds move together over time: a rise in short-term interest rates indicates that short-term interest rates will, on average, be higher in the future, and the first term in Equation 3 then implies that long-term interest rates will rise along with them.

They also explain why yield curves tend to have an especially steep upward slope when short-term interest rates are low and to be inverted when short-term rates are high (fact 2). Because investors generally expect short-term interest rates to rise to some normal level when they are low, the average of future expected short-term rates will be high relative to the current short-term rate. With the additional boost of a positive liquidity premium, long-term interest rates will be substantially above current short-term rates, and the yield curve would then have a steep upward slope. Conversely, if short-term rates are high, people usually expect them to come back down. Long-term rates would then drop below short-term rates because the average of expected future short-term rates would be so far below current short-term rates that despite positive liquidity premiums, the yield curve would slope downward.

The liquidity premium and preferred habitat theories explain fact 3, that yield curves typically slope upward, by recognizing that the liquidity premium rises with a bond's maturity because of investors' preferences for short-term bonds. Even if

short-term interest rates are expected to stay the same on average in the future, long-term interest rates will be above short-term interest rates, and yield curves will typically slope upward.

How can the liquidity premium and preferred habitat theories explain the occasional appearance of inverted yield curves if the liquidity premium is positive? It must be that at times short-term interest rates are expected to fall so much in the future that the average of the expected short-term rates is well below the current short-term rate. Even when the positive liquidity premium is added to this average, the resulting long-term rate will still be below the current short-term interest rate.

As our discussion indicates, a particularly attractive feature of the liquidity premium and preferred habitat theories is that they tell you what the market is predicting about future short-term interest rates just from the slope of the yield curve. A steeply rising yield curve, as in panel (a) of Figure 6-6, indicates that short-term interest rates are expected to rise in the future. A moderately steep yield curve, as in panel (b), indicates that short-term interest rates are not expected to rise or fall much in the future. A flat yield curve, as in panel (c), indicates that short-term rates are expected to fall moderately in the future. Finally,

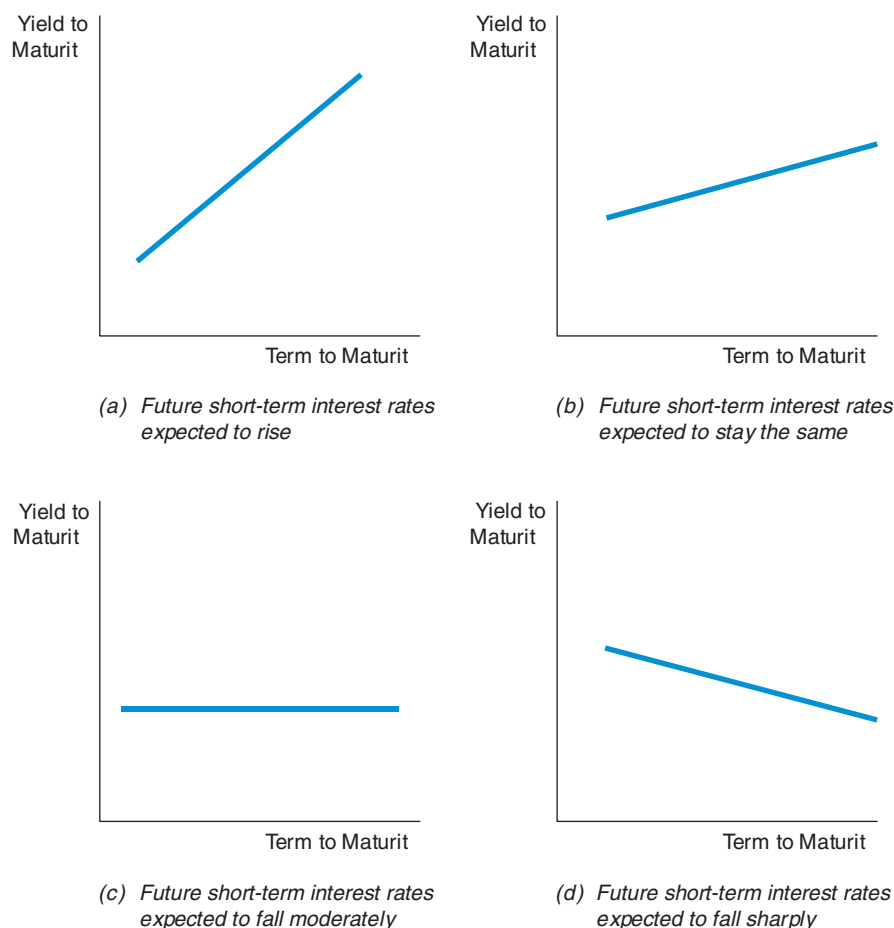


FIG RE 6-6 Yield Curves and the Market's Expectations of Future Short-Term Interest Rates According to the Liquidity Premium and Preferred Habitat Theories

an inverted yield curve, as in panel (d), indicates that short-term interest rates are expected to fall sharply in the future.

The Predictive Power of the Yield Curve

People often think that the slope of the yield curve can be used to forecast future short-term interest rates. The yield curve has this practical use only if it is determined by the expectations theory of the term structure that views long-term interest rates as equalling the average of future short-term interest rates. If, however, there are liquidity (term) premiums in the term structure, then it will be difficult to extract a reliable forecast of future short-term interest rates without good measures of these premiums.

In the 1980s, researchers examining the term structure of interest rates questioned whether the slope of the yield curve provides information about movements of future short-term interest rates.⁴ They found that the spread between long- and short-term interest rates does not always help predict future short-term interest rates, a finding that may stem from substantial fluctuations in the liquidity (term) premium for long-term bonds. More recent research using more discriminating tests now favours a different view. It shows that the term structure contains quite a bit of information for the very short run (over the next several months) and the long run (over several years) but is unreliable at predicting movements in interest rates over the intermediate term (the time in between).⁵ Research also finds that the yield curve helps forecast future inflation and business cycles (see the FYI box).

Summary

The liquidity premium and preferred habitat theories are the most widely accepted theories of the term structure of interest rates because they explain the major empirical facts about the term structure so well. They combine the features of both the expectations theory and the segmented markets theory by asserting that a long-term interest rate will be the sum of a liquidity (term) premium and the average of the short-term interest rates that are expected to occur over the life of the bond.

The liquidity premium and preferred habitat theories explain the following facts: (1) Interest rates on bonds of different maturities tend to move together over time, (2) yield curves usually slope upward, and (3) when short-term interest rates are low, yield curves are more likely to have a steep upward slope, whereas when short-term interest rates are high, yield curves are more likely to be inverted.

The theories also help us predict the movement of short-term interest rates in the future. A steep upward slope of the yield curve means that short-term rates are expected to rise, a mild upward slope means that short-term rates are expected to

⁴ Robert J. Shiller, John Y. Campbell, and Kermit L. Schoenholtz, "Forward Rates and Future Policy: Interpreting the Term Structure of Interest Rates," *Brookings Papers on Economic Activity* 1 (1983): 173–217; N. Gregory Mankiw and Lawrence H. Summers, "Do Long-Term Interest Rates Overreact to Short-Term Interest Rates?" *Brookings Papers on Economic Activity* 1 (1984): 223–242.

⁵ Eugene Fama, "The Information in the Term Structure," *Journal of Financial Economics* 13 (1984): 509–528; Eugene Fama and Robert Bliss, "The Information in Long-Maturity Forward Rates," *American Economic Review* 77 (1987): 680–692; John Y. Campbell and Robert J. Shiller, "Cointegration and Tests of the Present Value Models," *Journal of Political Economy* 95 (1987): 1062–1088; John Y. Campbell and Robert J. Shiller, "Yield Spreads and Interest Rate Movements: A Bird's Eye View," *Review of Economic Studies* 58 (1991): 495–514.

FYI

The Yield Curve as a Forecasting Tool for Inflation and the Business Cycle

Because the yield curve contains information about future expected interest rates, it should also have the capacity to help forecast inflation and real output fluctuations. To see this, recall from Chapter 5 that rising interest rates are associated with economic booms and falling interest rates with recessions. When the yield curve is either flat or negatively sloped, it suggests that future short-term interest rates will be falling and therefore that the economy is more likely to enter a recession. Indeed, the yield curve is found to be an accurate predictor of the business cycle.¹

In Chapter 4, we also learned that a nominal interest rate is composed of a real

interest rate and expected inflation, implying that the yield curve contains information not only about the future path of nominal interest rates, but about future inflation as well. A steep yield curve does predict a future rise in inflation, while a flat or negatively sloped yield curve forecasts a future fall in inflation.²

The ability of the yield curve to forecast business cycles and inflation is one reason why the slope of the yield curve is in the toolkit of many economic forecasters and is often viewed as a useful indicator of the stance of monetary policy, with a steep yield curve indicating loose policy and a flat or negatively sloped yield curve indicating tight policy.

¹ E.g., see Arturo Estrella and Frederic S. Mishkin, "Predicting U.S. Recessions: Financial Variables As Leading Indicators," *Review of Economics and Statistics* 80 (February 1998): 45–61.

² Frederic S. Mishkin, "What Does the Term Structure Tell Us About Future Inflation?" *Journal of Monetary Economics* 25 (January 1990): 77–95; and "The Information in the Longer-Maturity Term Structure About Future Inflation," *Quarterly Journal of Economics* 55 (August 1990): 815–28.

remain the same, a flat slope means that short-term rates are expected to fall moderately, and an inverted yield curve means that short-term rates are expected to fall sharply.

APPLICATION

Interpreting Yield Curves, 1990–2009

Figure 6-7 illustrates several yield curves that have appeared for Canadian government bonds in recent years. What do these yield curves tell us about the public's expectations of future movements of short-term interest rates?

The inverted yield curve that occurred in January 1990 indicated that short-term interest rates were expected to decline sharply in the future. In order for longer-term interest rates with their positive liquidity premium to be well below the short-term interest rate, short-term interest rates must be expected to decline so sharply that their average is far below the current short-term rate. Indeed, the public's expectations of sharply lower short-term interest rates evident in the yield curve were realized soon after January 1990; by March 1991, three-month treasury bill rates had declined from over 12% to less than 9%.

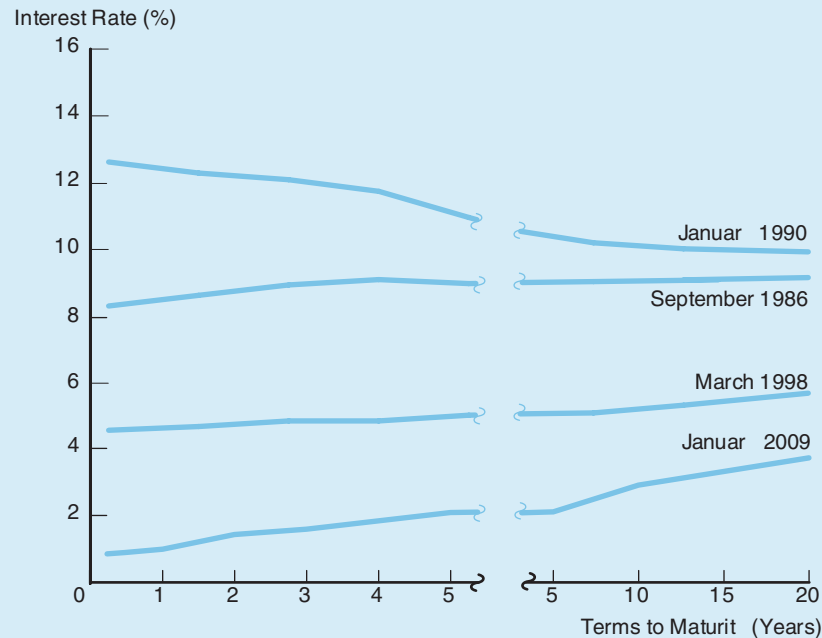


FIG RE 12-7 Yield Curves for Government of Canada Bonds, 1986–2009

Source: Statistics Canada CANSIM II Series V122531, V122532, V122533, V122538, V122539, V122540, V122543, and V122544, and the authors' calculations.

The upward-sloping yield curve in January 2009 indicates that short-term interest rates would climb in the future. The long-term interest rate is above the short-term interest rate when short-term interest rates are expected to rise because their average plus the liquidity premium will be above the current short-term rate. The moderately upward-sloping yield curve in September 1986 indicates that short-term interest rates were expected neither to rise nor to fall in the near future. In this case, their average remains the same as the current short-term rate, and the positive liquidity premium for longer-term bonds explains the moderate upward slope of the yield curve.

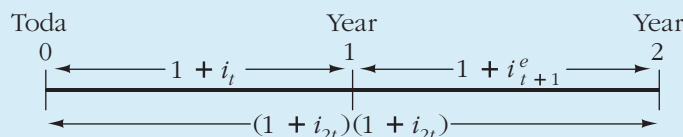
APPLICATION

Using the Term Structure to Forecast Interest Rates

Interest-rate forecasts are extremely important to managers of financial institutions because future changes in interest rates have a significant impact on the profitability of their institutions. Furthermore, interest-rate forecasts are needed when managers of financial institutions have to set interest rates on loans that are promised to customers in the future. Our discussion of the term structure of interest rates has indicated that the slope of the yield curve provides general information about the market's prediction of the future path of interest rates. For example, a steeply

upward-sloping yield curve indicates that short-term interest rates are predicted to rise in the future, and a downward-sloping yield curve indicates that short-term interest rates are predicted to fall. However, a financial institution manager needs much more specific information on interest-rate forecasts than this. Here we show how the manager of a financial institution can generate specific forecasts of interest rates using the term structure.

To see how this is done, let's start the analysis using the approach we took in developing the expectations theory. Recall that because bonds of different maturities are considered perfect substitutes, we assumed that the expected return over two periods from investing \$1 in a two-period bond, which is $(1 + i_2)(1 + i_2) - 1$, must equal the expected return from investing \$1 in one-period bonds, which is $(1 + i)(1 + i_{+1}^e) - 1$. This is shown graphically as follows:



In other words,

$$(1 + i)(1 + i_{+1}^e) - 1 = (1 + i_2)(1 + i_2) - 1$$

Through some tedious algebra we can solve for i_{+1}^e :

$$i_{+1}^e = \frac{(1 + i_2)^2}{1 + i} - 1 \quad (4)$$

This measure of i_{+1}^e is called the **forward rate** because it is the one-period interest rate that the pure expectations theory of the term structure indicates is expected to prevail one period in the future. To differentiate forward rates derived from the term structure from actual interest rates that are observed at time t , we call these observed interest rates **spot rates**.

Going back to the Application Expectations Theory and the Yield Curve (p. 124), which we used to discuss the expectations theory earlier in this chapter, at time t the one-year interest rate is 5% and the two-year rate is 5.5%. Plugging these numbers into Equation 4 yields the following estimate of the forward rate one period in the future:

$$i_{+1}^e = \frac{(1 + 0.055)^2}{1 + 0.05} - 1 = 0.06 = 6\%$$

Not surprisingly, this 6% forward rate is identical to the expected one-year interest rate one year in the future. This is exactly what we should find, as our calculation here is just another way of looking at the pure expectations theory.

We can also compare holding the three-year bond against holding a sequence of one-year bonds, which reveals the following relationship:

$$(1 + i)(1 + i_{+1}^e)(1 + i_{+2}^e) - 1 = (1 + i_3)(1 + i_3)(1 + i_3) - 1$$

and plugging in the estimate for i_{+1}^e derived in Equation 4, we can solve for i_{+2}^e :

$$i_{+2}^e = \frac{(1 + i_3)^3}{(1 + i_2)^2} - 1$$

Continuing with these calculations, we obtain the general solution for the forward rate n periods into the future:

$$i_{t+n}^e = \frac{(1 + i_{n+1t})^{n+1}}{(1 + i_{nt})^n} - 1 \quad (5)$$

Our discussion indicated that the expectations theory is not entirely satisfactory because investors must be compensated with liquidity premiums to induce them to hold longer-term bonds. Hence, we need to modify our analysis, as we did when discussing the liquidity premium theory, by allowing for these liquidity premiums in estimating predictions of future interest rates.

Recall from the discussion of those theories that because investors prefer to hold short-term rather than long-term bonds, the n -period interest rate differs from that indicated by the pure expectations theory by a liquidity premium of ℓ_{nt} . So to allow for liquidity premiums, we need merely subtract ℓ_{nt} from i_{nt} in our formula to derive i_{t+n}^e :

$$i_{t+n}^e = \frac{(1 + i_{n+1t} - \ell_{n+1t})^{n+1}}{(1 + i_{nt} - \ell_{nt})^n} - 1 \quad (6)$$

This measure of i_{t+n}^e is referred to, naturally enough, as the *adjusted forward-rate forecast*.

In the case of i_{t+1}^e , Equation 6 produces the following estimate:

$$i_{t+1}^e = \frac{(1 + i_{2t} - \ell_{2t})^2}{1 + i_{1t}} - 1$$

Using the example from the Liquidity Premium Theory Application on page 128, at time t the ℓ_{2t} liquidity premium is 0.25%, $\ell_{1t} = 0$, the one-year interest rate is 5%, and the two-year interest rate is 5.75%. Plugging these numbers into our equation yields the following adjusted forward-rate forecast for one period in the future:

$$i_{t+1}^e = \frac{(1 + 0.0575 - 0.0025)^2}{1 + 0.05} - 1 = 0.06 = 6\%$$

which is the same as the expected interest rate used in the Application on expectations theory, as it should be.

Our analysis of the term structure thus provides managers of financial institutions with a fairly straightforward procedure for producing interest-rate forecasts. First they need to estimate ℓ_{nt} , the values of the liquidity premiums for various n . Then they need merely apply the formula in Equation 6 to derive the market's forecasts of future interest rates.

APPLICATION

Forward Rate

A customer asks a bank if it would be willing to commit to making the customer a one-year loan at an interest rate of 8% one year from now. To compensate for the costs of making the loan, the bank needs to charge one percentage point more than the expected interest rate on a Canada bond with the same maturity if it is to make a profit. If the bank manager estimates the liquidity premium to be 0.4%, and the one-year Canada bond rate is 6% and the two-year bond rate is 7%, should the manager be willing to make the commitment?

Solution

The bank manager is unable to make the loan because at an interest rate of 8%, the loan is likely to be unprofitable to the bank.

$$i_{n+1}^e = \frac{(1 + i_{n+1} - \ell_{n+1})^{n+1}}{(1 + i_n - \ell_n)^n} - 1$$

where

$$\begin{aligned} i_{n+1} &= \text{two-year bond rate} = 0.07 \\ \ell_{n+1} &= \text{liquidity premium} = 0.004 \\ i_n &= \text{one-year bond rate} = 0.06 \\ \ell_1 &= \text{liquidity premium} = 0 \\ n &= \text{number of years} = 1 \end{aligned}$$

Thus

$$i_{n+1}^e = \frac{(1 + 0.07 - 0.004)^2}{1 + 0.06} - 1 = 0.072 = 7.2\%$$

The market's forecast of the one-year Canada bond rate one year in the future is therefore 7.2%. Adding the 1% necessary to make a profit on the one-year loan means that the loan is expected to be profitable only if it has an interest rate of 8.2% or higher.

SUMMARY

1. Bonds with the same maturity will have different interest rates because of three factors: default risk, liquidity, and tax considerations. The greater a bond's default risk, the higher its interest rate relative to other bonds; the greater a bond's liquidity, the lower its interest rate; and bonds with tax-exempt status will have lower interest rates than they otherwise would. The relationship among interest rates on bonds with the same maturity that arises because of these three factors is known as the risk structure of interest rates.
2. Four theories of the term structure provide explanations of how interest rates on bonds with different terms to maturity are related. The expectations theory views long-term interest rates as equalling the average of future short-term interest rates expected to occur over the life of the bond; by contrast, the segmented markets theory treats the determination of interest rates for each bond's maturity as the outcome of supply and demand in each market in isolation. Neither of these theories by itself can explain the fact that interest rates on bonds of different maturities move together over time and that yield curves usually slope upward.
3. The liquidity premium and preferred habitat theories combine the features of the other two theories and by so doing are able to explain the facts just mentioned. They view long-term interest rates as equalling the average of future short-term interest rates expected to

occur over the life of the bond plus a liquidity premium. These theories allow us to infer the market's expectations about the movement of future short-term interest rates from the yield curve. A steeply upward-sloping curve indicates that future short-term rates are expected to rise, a mildly upward-sloping curve indi-

cates that short-term rates are expected to stay the same, a flat curve indicates that short-term rates are expected to decline slightly, and an inverted yield curve indicates that a substantial decline in short-term rates is expected in the future.

KEY TERMS

basis point, p. 117

credit-rating agencies, p. 115

default-free bonds, p. 114

expectations theory, p. 121

fallen angels, p. 116

forward rate, p. 134

inverted yield curve, p. 120

junk bonds, p. 116

liquidity premium theory,
p. 127

preferred habitat theory, p. 127

risk of default, p. 114

risk premium, p. 114

risk structure of interest rates,
p. 113

segmented markets theory, p. 126

spot rate, p. 134

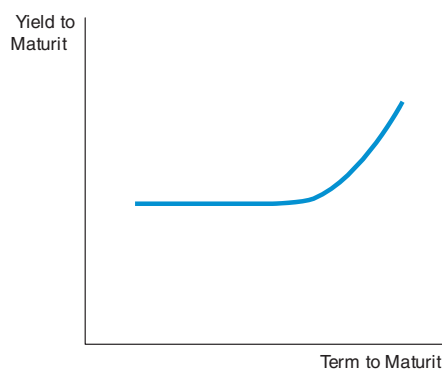
term structure of interest rates,
p. 113

yield curve, p. 119

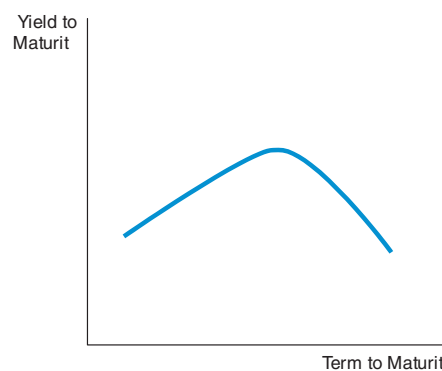
QUESTIONS

You will find the answers to the questions marked with an asterisk in the Textbook Resources section of your MyEconLab.

1. Which should have the higher risk premium on its interest rates, a corporate bond with an S&P BBB rating or a corporate bond with a C rating? Why?
- *2. Why do Canadian treasury bills have lower interest rates than large-denomination negotiable bank CDs?
3. Risk premiums on corporate bonds are usually anti-cyclical; that is, they decrease during business cycle expansions and increase during recessions. Why is this so?
- *4. "If bonds of different maturities are close substitutes, their interest rates are more likely to move together." Is this statement true, false, or uncertain? Explain your answer.
5. If yield curves, on average, were flat, what would this say about the liquidity (term) premiums in the term structure? Would you be more or less willing to accept the expectations theory?
- *6. If a yield curve looks like the one shown in (a), what is the market predicting about the movement of future short-term interest rates? What might the yield curve indicate about the market's predictions about the inflation rate in the future?



(a)



(b)

7. If a yield curve looks like the one shown in (b), what is the market predicting about the movement of future short-term interest rates? What might the yield curve indicate about the market's predictions about the inflation rate in the future?

- *8. What are the financial implications of a firm with a high default risk?

Predicting the Future

9. Predict what will happen to interest rates on a corporation's bonds if the federal government guarantees today that it will pay creditors if the corporation goes bankrupt in the future. What will happen to the interest rates on Canada bonds?
- *10. Predict what would happen to the risk premiums on corporate bonds if brokerage commissions were lowered in the corporate bond market.
11. Predict what would happen to yield spreads in response to the following macroeconomic events: recession, high inflation, and stock market increase.
- *12. If the yield curve suddenly becomes steeper, how would you revise your predictions of interest rates in the future?
13. If expectations of future short-term interest rates suddenly fall, what would happen to the slope of the yield curve?

QUANTITATIVE PROBLEMS

- *1. Assuming that the expectations theory is the correct theory of the term structure, calculate the interest rates in the term structure for maturities of one to five years, and plot the resulting yield curves for the following series of one-year interest rates over the next five years:

(a) 5%, 7%, 7%, 7%, 7%

(b) 5%, 4%, 4%, 4%, 4%

How would your yield curves change if people preferred shorter-term bonds to longer-term bonds?

2. Assuming that the expectations theory is the correct theory of the term structure, calculate the interest rates in the term structure for maturities of one to five years, and plot the resulting yield curves for the following path of one-year interest rates over the next five years:

(a) 5%, 6%, 7%, 6%, 5%

(b) 5%, 4%, 3%, 4%, 5%

How would your yield curves change if people preferred shorter-term bonds to longer-term bonds?

- *3. Rates on one-year T-bills over the next four years are expected to be 3%, 4%, 5%, and 5.5%. If four-year Canada bonds are yielding 4.5%, what is the liquidity premium on this bond?
4. Suppose that you are forecasting one-year T-bill rates as follows:

Year	1-year rate (%)
1	4.25
2	5.15
3	5.50
4	6.25
5	7.10

You have a liquidity premium of 0.25% for the next year and 0.50% thereafter. Would you be willing to purchase a four-year Canada bond at a 5.75% interest rate?

CANSIM Questions

5. Get the monthly data from 1978 to 2006 on the three-month T-bill rate (CANSIM series V122531), the interest rate on long-term corporate bonds (series V122518), and the interest rate on long-term Canada bonds (series V122544) from the Textbook Resources area of the MyEconLab.
- Present a time series plot of these interest rate series and comment on their long-run movements.
 - Calculate the mean and standard deviation as well as the maximum and minimum values for each series over the sample period.
 - Which were the worst and best years in terms of interest rates?
6. Get the monthly data from 1978 to 2006 on long-term Canada bonds (CANSIM series V122544), the interest rate on long-term provincial bonds (series V122517), and the interest rate on long-term corporate bonds (series V122518) from the Textbook Resources area of the MyEconLab.
- Present a time series plot of these interest rates and comment on their long-term movements.
 - Which series exhibit the strongest correlations? The weakest? Do the correlation patterns you identified here manifest in the graphical representation of the series?
 - Compare the contemporaneous correlations over the whole period with those in the 1960s, 1970s, 1980s, 1990s, and 2000s.
 - Calculate the corporate-Canada spread and the corporate-provincials spread and plot these series.
 - Continuing from (d), comment on the time paths of the risk premiums.

WEB EXERCISES

1. Go to **www.bloomberg.com** and click on “Market Data” and then “Rates & Bonds” to find information about the yield curve in Australia, Brazil, Germany, Hong Kong, Japan, the United Kingdom, and the United States. What does each of these yield curves tell us about the public’s expectations of future movements of short-term interest rates?
2. Investment companies attempt to explain to investors the nature of the risk the investor incurs when buying shares in their mutual funds. For example, Vanguard (a U.S. company) carefully explains interest rate risk and offers alternative funds with different interest rate risks. Go to **http://flagship.vanguard.com/VGApp/hnw/FundsStocksOverview**.
 - a. Select the bond fund you would recommend to an investor who has very low tolerance for risk and a short investment horizon. Justify your answer.
 - b. Select the bond fund you would recommend to an investor who has very high tolerance for risk and a long investment horizon. Justify your answer.


myeconlab

Be sure to visit the MyEconLab website at **www.myeconlab.com**. This online homework and tutorial system puts you in control of your own learning with study and practice tools directly correlated to this chapter content.

On the MyEconLab website you will find the following mini-case for this chapter:

Mini-Case 6.1: Yield Curve Hypotheses and the Effects of Economic Events

CHAPTER

13

The United States in a Global Economy

Learning Objectives

After studying Chapter 1, students will be able to:

- Explain how economists measure international economic integration.
- List the three types of evidence to support the idea that trade supports economic growth.
- Discuss the differences in international economic integration at the end of the nineteenth century and the current era.
- Describe the major themes of international economics.

Introduction: International Economic Integration

In August of 2007, a crisis erupted in the housing sector of the United States. At the time, few people realized that the subprime mortgage crisis would become a demonstration of international economic integration or that it would push the world economy to the brink of collapse. The crisis grew through the remainder of 2007 and into 2008, so that by the summer nearly all high-income economies were in deep distress. Contagion from the crisis spread like an epidemic as banks and other financial firms collapsed and solvent firms stopped lending. The scarcity of credit caused difficulties for businesses that could not find financing for their day-to-day operations while, at the same time, consumers cut back on their spending and businesses cut back on new investment. By the end of 2008, economies around the world were in recession, with the notable exceptions of China, India, and the major oil producers.

This episode is the most dramatic instance since the Great Depression of the 1930s of a crisis leading to severe economic recession in many countries around the world. It is, however, only one of several recent examples of crises spilling across national borders. The Russian Crisis of 1998–99, the Asian Crisis of 1997–98, the Mexican Crisis of 1994–95, the Latin American Debt Crisis of 1982–89, and a number of others caused major damage to financial systems, businesses, and households, both in the places where they originated and in many other countries.

The international integration of national economies has brought many benefits to nations across the globe, including technological innovation, less expensive products,

and greater investment in regions where local capital is scarce, to name a few. But it has also made countries vulnerable to economic problems that have become more easily transmitted from one place to another. Given that the benefits and costs of international economic integration are surrounded by controversy, it is worth clarifying what we mean by the term *international economic integration*, or *globalization in the economic sphere*. To help us understand these forces better, a historical perspective is also useful.

Elements of International Economic Integration

Most people would agree that the major economies of the world are more integrated than at any time in history. Given our instantaneous communications, modern transportation, and relatively open trading systems, most goods can move from one country to another without major obstacles and at relatively low cost. For example, most cars today are made in fifteen or more countries after you consider where each part is made, where the advertising originates, who does the accounting, and who transports the components and the final product. Nevertheless, the proposition that today's economies are more integrated than at any other time in history is not simple to demonstrate. It is clear that our current wave of economic integration began in the 1950s, with the reduction of trade barriers after World War II. In the 1970s, many countries began to encourage financial integration by increasing the openness of their capital markets. The advent of the Internet in the 1990s, along with the other elements of the telecommunications revolution, pushed economic integration to new levels as multinational firms developed international production networks and markets became ever more tightly linked.

Today's global economy is not the first instance of a dramatic growth in economic ties between nations, however, as there was another important period between approximately 1870 and 1913. New technologies such as transatlantic cables, steam-powered ships, railroads, and many others led the way, much as they do today. For example, when the first permanent transatlantic cable was completed in 1866, the time it took for a New York businessperson to complete a financial transaction in London fell from approximately three weeks to one day, and by 1914 it had fallen to one minute as radio telephony became possible.

We have mostly forgotten about this earlier period of economic integration, and that makes it easier to overestimate integration today. Instantaneous communications and rapid transportation, together with the easy availability of foreign products, often cause us to lose sight of the fact that most of what we buy and sell never makes it out of our local or national markets. We rarely pause to think that haircuts, restaurant meals, gardens, health care, education, utilities, and many other goods and services are partially or wholly domestic products. In the United States, for example, about 82.3 percent of goods and services are produced domestically, with imports (17.7 percent) making up the remainder of what we consume (2011).

By comparison, in 1890 the United States made about 92 percent of its goods and services, a larger share than today, but not radically different.

The question as to whether we are more economically integrated today or in some period in the past is not only academic. Between the onset of World War I in 1914 and the end of World War II in 1945, the world economy suffered a series of human-made catastrophes that de-integrated national economies. Two world wars and a global depression caused most countries to close their borders to foreign goods, foreign capital, and foreign people. Since the end of World War II, many of the economic linkages between nations have served to repair the damage done during the first half of the twentieth century, but there is no reason to think that events might not cause a similar decoupling in the future.

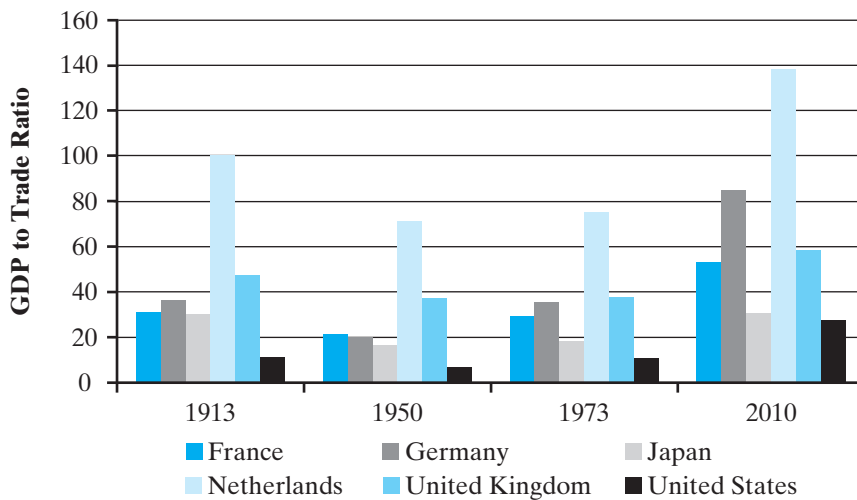
Understanding international economic integration requires us to define what we mean by the term. Economists usually point to four criteria or measures for judging the degree of integration. These are trade flows, capital flows, people flows, and the similarity of prices in separate markets. The first three points are relatively self-explanatory, while the similarity of prices refers to the fact that integrated economies have price differences that are relatively small and are due mainly to differences in transportation costs. Goods that can move freely from a low-cost to a high-cost region should experience price convergence as goods move from where they are plentiful and cheap to where they are relatively scarcer and more expensive. All of these indicators—trade flows, factor (labor and capital) movements, and similarity of prices—are measures of the degree of international economic integration.

The Growth of World Trade

Since the end of World War II, world trade has grown much faster than world output. One way to show this is to estimate the ratio of exports by all countries to total production by all countries. In 1950, total world exports—which are the same as world imports—are estimated to have been 5.5 percent of world **gross domestic product (GDP)**, a measure of total production. Fifty-five years later, in 2005, they were 20.5 percent of world GDP, nearly four times more important relative to the size of the world economy. One measure of the importance of international trade in a nation's economy is the sum of exports plus imports, divided by the GDP. Specifically, it is the value of all final goods and services produced inside a nation during some period, usually one year. The **trade-to-GDP ratio** is represented as follows:

$$\text{Trade to GDP ratio} = (\text{Exports} + \text{Imports}) \div \text{GDP}$$

The ratio does not tell us about a country's trade policies and countries with higher ratios do not necessarily have lower barriers to trade, although that is one possibility. In general, large countries are less dependent on international trade because their firms can reach an optimal production size without having

FIGURE 13.1 Trade-to-GDP Ratios for Six Countries, 1913–2010

Source: Maddison, A. (1991). "Dynamic Forces in Capitalist Development" and The World Trade Organization, "Statistics Database: Trade Profiles." Note that 2010 is an average of 2008–2010.

to sell to foreign markets. Consequently, smaller countries tend to have higher ratios of trade-to-GDP.

Figure 1.1 shows the trade-to-GDP ratio for six countries between 1913 and 2010. The decline in trade between the onset of World War I and 1950 is clearly visible in each country, as is the subsequent increase after 1950. Another pattern shown in Figure 1.1 is the smaller ratios for the United States and Japan, which have the largest populations, and the much higher ratio for the Netherlands, which has the smallest population in the sample. In general, smaller countries trade more than larger ones since they cannot efficiently produce as wide a range of goods and must depend on trade to a greater extent. For example, if the Netherlands were to produce autos solely for its own market, it would lack economies of scale and could not produce at a competitive cost, whereas the U.S. market can absorb a large share of U.S. output. Hence, the trade-to-GDP ratio measures the relative importance of international trade in a nation's economy, but it does not provide any direct information about trade policy or trade barriers.

Figure 1.1 gives a historical overview of the decline and subsequent return of international trade after World War II, but it obscures important changes in the composition of trade flows from early in the twentieth century to those at the end of the century. Before World War I most trade consisted of agricultural commodities and raw materials, while current trade is primarily manufactured consumer goods and producer goods (machinery and equipment). Consequently, today's manufacturers are much more exposed to international com-

petition than was the case in 1900. In addition, much of the growth of world trade since 1950 has been accomplished by multinational corporations. With production sites in multiple countries and inputs that pass back and forth between affiliates, multinational corporations have become dramatically important. This trend has been supported and encouraged by the telecommunications revolution and transportation improvements that have lowered the costs of coordinating operations physically separated by oceans and continents. And finally, it has also become possible to coordinate service operations such as accounting and data processing from a great distance. In sum, trade today is qualitatively different than in 1913, and the growth of the trade-to-GDP ratio since 1950 does not tell the whole story.

Capital and Labor Mobility

In addition to exports and imports, factor movements also are an indicator of economic integration. As national economies become more interdependent, labor and capital should move more easily across international boundaries. Labor, however, is less mobile internationally than it was in 1900. Consider, for example, that in 1890 approximately 14.5 percent of the U.S. population was foreign born, while in 2010, the figure was 12.9 percent. In 1900, many nations had open door immigration policies, and passport controls, immigration visas, and work permits were exceptions rather than rules. The movement of people was severely restricted by the two world wars and the Great Depression of the 1930s. In the 1920s, during the interwar period, the United States sharply restricted immigration with policies that lasted until the 1960s, when changes in immigration laws once again encouraged foreigners to migrate to the United States.

On the capital side, measurement is more difficult, since there are several ways to measure capital flows. The most basic distinction is between flows of financial capital representing paper assets such as stocks, bonds, currencies, and bank accounts, and flows of capital representing physical assets such as real estate, factories, and businesses. The latter type of capital flow is called **foreign direct investment (FDI)**. To some extent, the distinction between the two types of capital flows is immaterial because both represent shifts in wealth across national boundaries and both make one nation's savings available to another.

When we compare international capital flows today to a century ago, there are two points to keep in mind. First, savings and investment are highly correlated. That is, countries with high savings tend to have high rates of investment, and low savings is correlated with low investment. If there were a single world market in which capital flowed freely and easily, this would not necessarily be the case. Capital would flow from countries with abundant savings and capital to countries with low savings and capital, where it would find its highest returns. Second, a variety of technological improvements increased capital flows in the 1800s, as they are doing today. Transoceanic cables and radio telephony have already been

mentioned, but capital flows also increased in the late 1800s because there were new investment opportunities such as national railroad networks and other infrastructure, both at home and abroad.

If we compare the size of capital flows today to the previous era of globalization, flows today are much larger but mainly because economies are larger. Relative to the size of economies, the differences are not great and may even favor the 1870 to 1913 period, depending on what is measured. Great Britain routinely invested 9 percent of its GDP abroad in the decades before 1913, and France, Germany, and the Netherlands were as high at times. For significant periods, Canada, Australia, and Argentina borrowed amounts that exceeded 10 percent of their GDP, a level of borrowing that sends up danger signals in the world economy today. In other words, it is hard to make the argument that national economies have a historically unprecedented level of international capital flows today.

While the relative quantity of capital flows today may not be that much different for many countries, there are some important qualitative differences. First, there are many more financial instruments available now than there were a century ago. These range from relatively mundane stocks and bonds to relatively exotic instruments such as derivatives, currency swaps, and others. By contrast, at the turn of the twentieth century there were many fewer companies listed on the world's stock exchanges and most international financial transactions involved the buying and selling of bonds.

A second difference today is the role of foreign exchange transactions. In 1900, countries had fixed exchange rates and firms in international trade or finance had less day-to-day risk from a sudden change in the value of a foreign currency. Many firms today spend significant resources to protect themselves from sudden shifts in currency values. Consequently, buying and selling assets denominated in foreign currencies is the largest component of international capital movements. For example, according to the Bank for International Settlements in Geneva, Switzerland, *daily* foreign exchange transactions in 2010 were equal to \$3.98 *trillion*. In 1973, at the end of the last era of fixed exchange rates, they were \$15 billion.

The third major difference in capital flows is that the costs of foreign financial transactions have fallen significantly. Economists refer to the costs of obtaining market information, negotiating an agreement, and enforcing the agreement as **transaction costs**. They are an important part of any business's costs, whether it is a purely domestic enterprise or a company involved in foreign markets. Due to sheer distance, as well as differences in culture, laws, and languages, transaction costs are often higher in international markets than in domestic ones. Today's lower transaction costs for foreign investment mean that it is less expensive to move capital across international boundaries.

The volatile movement of financial capital across international boundaries is often mistakenly regarded as a new feature of the international economy. Speculative excesses and overinvestment, followed by capital flight and bankruptcies, have occurred throughout the modern era, going back at least to the 1600s and

probably earlier. U.S. and world history show a number of such cases. Financial crises are not a new phenomenon, nor have we learned how to avoid them—a fact driven home by the recent subprime mortgage crisis.

Features of Contemporary International Economic Relations

While international economic integration has been rapid, it does not appear to be historically unprecedented. The trade-to-GDP ratio is about 50 percent higher in the U.S. economy than it was in 1890, and manufacturers and service providers are more exposed to international forces. Labor is less mobile than in 1900 due to passport controls and work permits, but capital is more mobile and encompasses a larger variety of financial forms. Prices in many U.S. and foreign markets tend to be similar, although there are still significant differences. In quantitative terms, the differences between today and a hundred years ago may not be as great as many people imagine, but qualitatively, a number of additional features of the world economy separate the first decade of the twenty-first century from the first decade of the twentieth.

Deeper Integration High-income countries have low barriers to imports of manufactured goods. There are some exceptions (processed foodstuffs and apparel), but as a general rule import **tariffs** (taxes on imports) and other barriers such as **quotas** (quantitative restrictions on imports) are much less restrictive than they were in the middle of the twentieth century. As trade barriers came down during the second half of the twentieth century, two other trends began to intensify economic integration between countries. First, lower trade barriers exposed the fact that most countries have domestic policies that are obstacles to international trade. National regulations governing labor, environmental, and consumer safety standards; rules governing investment location and performance; rules defining fair and unfair competition; rules on government “buy-national” programs; and government support policies for specific industries—all have little impact on trade until formal trade barriers start to fall and trade volume increases. These policies were not implemented to protect domestic industries from foreign competition, and as long as tariffs were high and trade flows were limited, they did not matter much to trade relations. Once tariffs fell, however, many forms of domestic policies began to be viewed as barriers to increased trade. Economists sometimes refer to the reduction of tariffs and the elimination of quotas as **shallow integration** and negotiations over domestic policies that impact international trade as **deep integration**. Deep integration is much more contentious than shallow integration and much more difficult to accomplish since it involves domestic policy changes that align a country with rules that are created abroad, or at least negotiated with foreign powers.

A second noticeable trend over the last few decades is that technologically complicated goods such as smart phones and automobiles are made of components produced in more than one country and, consequently, labels such as

“Made in China” or “Made in the USA” are less and less meaningful. Low tariffs along with innovations in transportation and communication technologies have enabled firms to locate production of the different components of a sophisticated product in different countries. For example, the hardware for a 3G iPhone is produced in Germany, Korea, Japan, and the United States, and then it is assembled in China. The most valuable share of the hardware is made in Japan, but no one thinks of this device as a Japanese phone. In this case, as in many others, it is not accurate to say the product is made in one particular country since the parts come from all over, and the product is the result of a multinational effort involving firms and workers from many different countries.

These two trends raise new issues that are shaping the world economy in the twenty-first century. The first trend, greater interest in the consequences of different domestic policies, makes trade negotiations more difficult and creates widespread discussion of labor, environmental, and other standards that may affect trade flows. The second trend, greater participation in the production of a single product by firms in multiple countries, leads to concerns about the impact of trade on national economies, employment, and working conditions. National and international dialogues on these issues are a key feature of international economics in the twenty-first century.

Multilateral Organizations At the end of World War II, the United States, Great Britain, and their allies created a number of international organizations to maintain international economic and political stability. Although the architects of these organizations could not envision the challenges and issues they would confront over the next fifty years, the organizations were given significant flexibility, and they continue to play an important and growing role in managing the issues of shallow and deeper integration.

The International Monetary Fund (IMF), the World Bank, the General Agreement on Tariffs and Trade (GATT), the United Nations (UN), the World Trade Organization (the WTO began operation in 1995, but grew out of the GATT), and a host of smaller organizations have broad international participation. They serve as forums for discussing and establishing rules, as mediators of disputes, and as organizers of actions to resolve problems. All of these organizations are controversial and have come under increasing fire from critics who charge that they promote unsustainable economic policies or that they protect the interests of wealthy countries. Others argue that they are unnecessary foreign entanglements that severely limit the scope for national action (Chapter 2 examines this issue in detail). These organizations are attempts to create internationally acceptable rules for trade and commerce and to deal with potential disputes before they spill across international borders; they are an entirely new element in the international economy.

Regional Trade Agreements Agreements between groups of nations are not new. Free-trade agreements and other forms of preferential trade have

existed throughout history. What is new is the significant increase in the number of **regional trade agreements (RTAs)** that have been signed in the last twenty years.

The formation of preferential trade agreements is controversial. Trade opponents dislike the provisions that expose more of the national economy to international competition, whereas some trade proponents dislike preferences that favor countries included in the agreement at the expense of countries outside the agreement. The North American Free Trade Agreement (NAFTA), the European Union (EU), the Mercado Común del Sur (MERCOSUR), and the Asia Pacific Economic Cooperation (APEC) are examples of RTAs, but more than 330 have been recorded by the World Trade Organization (2012).

Trade and Economic Growth

Many people are more than a little apprehensive about increased international economic integration. The list of potential problems is a long one. More trade may give consumers lower prices and greater choices, but it also means more competition for firms and workers. Capital flows make more funds available for investment purposes, but they also increase the risk of spreading financial crises internationally. Rising immigration means higher incomes for migrants and lower labor costs or a better pool of skills for firms, but it also means more competition in labor markets and, inevitably, greater social tensions. International organizations may help resolve disputes, but they may also reduce national sovereignty by putting pressure on countries to make operational changes. Free-trade agreements may increase trade flows, but again, that means more competition and more pressure on domestic workers and firms.

In general, economists remain firmly convinced that the benefits of trade outweigh the costs. There is disagreement over the best way to achieve different goals (for example, how to protect against the harmful effects of sudden flows of capital), but the general belief that openness to the world economy is a superior policy to closing off a country is quite strong. To support this stance, economists can point to the following kinds of evidence:

- Casual empirical evidence of historical experience
- Evidence based on economic models and deductive reasoning
- Evidence from statistical comparisons of countries

While none of these is conclusive by itself, together they provide solid support for the idea that open economies generally grow faster and prosper more than closed ones.

The historical evidence examines the experiences of countries that tried to isolate themselves from the world economy. There are the experiences of the 1930s, when most countries tried to protect themselves from world events by shutting out flows of goods, capital, and labor. This did not cause the Great

Depression of the 1930s, but it did worsen it, and ultimately it led to the misery and tragedy of World War II. There are also the parallel experiences of countries that were divided by war, with one side becoming closed to the world economy, and the other side open. Germany (East versus West), Korea (North versus South), and China (mainland China before the 1980s versus Taiwan and Hong Kong) are the best examples.

Economic theory generally supports these examples by suggesting the causal mechanisms that lead from trade to faster growth. Generally, the benefits of increased innovation, competitive pressure to raise productivity levels, and access to new technologies and ideas that are fostered by trade are positive factors. On the consumer side, trade provides a greater variety of goods and offers them at lower prices.

The statistical evidence of the benefits of more open economies comes from comparisons of large samples of countries over different periods. While the statistical tests of the relationship between trade policy and economic growth suffer from their own technical shortcomings, the results consistently show that more open economies grow faster. These results cannot be viewed as absolutely conclusive, but together with trade theory and the casual empirical evidence drawn from historical experiences, the available statistical analysis provides additional support for the notion that trade is usually beneficial.

Twelve Themes in International Economics

Each of the twelve themes discussed next are examined in the chapters that follow. These themes are overlapping, multidimensional, and often go beyond pure economics. International economic analysis cannot claim the final word, but it is hoped that it will provide you an analytically powerful and logically consistent approach for thinking about the issues raised by these themes.

The Gains from Trade and New Trade Theory

Why is international trade desirable? We have briefly addressed this issue, and we will consider additional points as we continue. Given that economic analysis clearly demonstrates that the benefits of international trade outweigh the costs, it is not surprising that virtually all economists generally support open markets and increased trade. The benefits of international trade were first analyzed in the late 1700s and are perhaps the oldest and strongest finding in all of economics. More recently, economists have begun to analyze returns to scale within firms and industries. Under the label “New Trade Theory,” economists have demonstrated a number of new sources of national welfare improvements due to international trade and added greater sophistication to our understanding of market structure and trade effects.

Wages, Jobs, and Protection

International trade raises national welfare, but it does not benefit every member of society. Workers in firms that cannot compete may be forced to find new jobs or take pay cuts. The fact that consumers pay less for the goods they buy, or that exporters hire more workers, may not help them. Increased awareness of the international economy has heightened the fears of people who feel vulnerable to change. They are concerned that wages in high-income countries must fall in order to compete with workers in low-wage countries, and that their jobs may be moved overseas. One of the key challenges for policymakers is to find the right mix of domestic policies so that the nation benefits from trade without creating a backlash from those individuals and industries that are hurt.

Trade Deficits

In 1980, a comprehensive measure of trade accounts in the United States showed that there was a slight surplus. Since then, only two of the following twenty-eight years had surpluses, and the sum of the deficits since 2000 is more than \$5.78 trillion (2001 through 2010). The United States was not the only country running deficits, but each year a country runs a deficit in its trade accounts, it must borrow from abroad, essentially selling a piece of its future output in order to obtain more goods and services today. As the United States and other countries borrowed, China, Germany, Japan, and oil producers like Saudi Arabia and Russia lent. These large imbalances in lending and borrowing played a key role in the crisis that began in 2007.

Regional Trade Agreements

As the world economy becomes more integrated, some regions are running ahead of the general trend. Western Europeans, for example, have eliminated many of the economic barriers separating their nations, and are creating a broad political and economic union. With implementation of NAFTA in 1994, the United States, Canada, and Mexico became a free-trade area while the largest countries in the Pacific Basin, including China, Japan, and the United States, have agreed to turn the Pacific region into another free-trade area by 2020. The three NAFTA countries each individually signed free trade agreements with most of Central America and the Dominican Republic, and the United States continues to negotiate with countries in South America and Asia. Since 2004, ten central and Eastern European countries have joined the EU, along with two small Mediterranean states. The ten members of the Association of South East Asian Nations (ASEAN) have moved to create a free-trade zone, and China has become an active participant in trade agreements, along with a number of other countries. The pros and cons of these and other agreements is an active area of economic interest and will be considered in several chapters.

The Resolution of Trade Conflicts

Commercial conflicts between nations cover a wide variety of issues and complaints. In one sense these conflicts are routine, as the WTO provides a formal dispute resolution procedure that has the assent of most of the world's nations. The WTO process does not cover all goods and services, however, nor does it say much about a large number of practices that some nations find objectionable. The ability of nations to resolve conflicts without resorting to protectionist measures is one key to maintaining a healthy international economic environment. Disputes can become acrimonious, so it is imperative that differences of opinion are not permitted to escalate into a wider disagreement. Trade wars are not real wars, but they are harmful nonetheless.

The Role of International Institutions

The organization with the greatest responsibility for resolving trade disagreements is the WTO. The WTO came into existence in 1995 and was an adaptation of the GATT, which was created shortly after World War II. Resolving trade disputes is only one of the new roles played by international organizations. Various organizations offer development support, technical economic advice, emergency loans in a crisis situation, and other services and assistance. These organizations perform services that were not offered before World War II (development support), or that were done by a single country (lending in a crisis)—usually the world's greatest military power. They exist today only through the mutual consent and cooperation of participating nations; without that cooperation, they would dissolve. Their abilities are limited, however. They cannot prevent crises, and they cannot make poor countries rich. They are also controversial and are viewed by some as tools of the United States or as a threat to national independence. They are very likely to grow in function, however, as many international problems cannot be solved by individual nations alone.

Exchange Rates and the Macroeconomy

Seventeen of the twenty-seven members of the EU have adopted the euro as a common currency, and several more are preparing to join them in spite of the euro crisis that began in 2011. Panama, El Salvador, and Ecuador use the U.S. dollar. Some members of the U.S. Congress and some economists think that China artificially manipulates its currency to gain commercial advantages, and China's leaders worry that the United States might let the dollar sink in value to depreciate its foreign debt. Exchange rate systems come in a variety of forms and link the domestic economy to the rest of the world. They can help protect a country against harmful developments outside its borders, but they can also magnify and transmit those developments to the domestic economy. Exchange rates play a key role in the international economy.

Financial Crises and Global Contagion

As international trade and investment barriers declined, and as new communications and transportation systems developed, increasing quantities of capital flowed across national borders. These flows were encouraged by financial innovation and a general spirit of deregulation that held sway in much of the world from the late 1970s forward. Capital flows brought many desirable things, such as investment, new technology, and higher consumption, but they also often outpaced our ability to monitor and supervise, and were frequently at the root of financial crises, including the severe global crisis that began in 2007. Economists are engaged in a broad discussion today, aimed at finding techniques for reducing the macroeconomic and financial volatility caused by capital flows without hampering the new investment and lending that they provide.

Capital Flows and the Debt of Developing Countries

In 1996, the World Bank and the IMF began a debt relief program for a group of forty-two countries labeled the *Highly Indebted Poor Countries (HIPC)*. Thirty-four of these countries are in Africa. At the same time, non-governmental groups and celebrities, such as Bono, began to lobby successfully for a reduction in the debts of poor countries and for changes in the lending policies of rich countries. In many parts of the world, problems of extreme poverty are compounded by large foreign debts that are unlikely to be repaid and often require a constant supply of new loans to pay interest on the old ones. The search for workable solutions is complicated in the borrowing countries by economic shocks, corruption, and unsustainable economic policies. Common problems in the lending countries include unwise loans to corrupt dictators and loans for some expensive and unnecessary goods sold by rich countries.

Latin America and the World Economy

In Latin America, the 1980s are known as the *Lost Decade*. High levels of debt, deep recessions, and hyperinflation caused the region to lose a decade of growth and development. In response, many countries embarked on a profound shift in their economic policies. They opened markets, allowed increased foreign investment, signed trade agreements, and ended a long period of relative isolation from the world economy. These policy changes became known as the Washington Consensus and helped to bring an end to the Lost Decade, but few economists think the policies were successful. Growth remained relatively low in many places, financial crises continued to undermine economic gains, and traditional issues of economic fairness were largely ignored. Latin American countries have developed a wide variety of new policies and experiments as they try to reduce poverty, generate prosperity, and provide opportunity for all their citizens.

Export-Led Growth in East Asia

Throughout the late 1980s and into the 1990s, it was hard to ignore the East Asian “miracle.” While some economists point out that it was not really a miracle—just a lot of hard work and sound economic policies—the growth rates of the “high-performance Asian economies” were unique in human history. Rates of growth of real GDP *per person* commonly reached 4 to 5 percent per year, with 6 to 8 percent not unusual. In 1997, an economic and financial crisis hit the region hard. Although there were lingering effects, by 2000 the economies of the region’s developing countries were growing at more than 7 percent a year. One of the dominant traits of the countries in East Asia is the extent to which they are outward looking and dependent on the growth of their manufactured exports.

The Integration of the BRICs into the World Economy

Between 1995 and 2001, world merchandise trade grew by a factor of 3.5. At the same time, the exports of the four **BRIC countries** (Brazil, Russia, India, and China) grew almost tenfold, so that their trade expanded from 5 percent of world trade to over 16 percent. The term *BRIC* is an artificial and somewhat fluid designation, but it is useful shorthand for middle-income countries that are experiencing rapid growth and having a significant impact on world trade and finance. China’s economy will likely surpass the U.S. economy to become the world’s largest economy in absolute size sometime early in the 2020s or possibly even sooner; and although the BRIC countries are considerably less well off than the richest countries of the world, their large populations and rapid growth are giving them an increasing voice in world affairs. In historical terms, the rise of these and other developing countries represents a return to a multipolar world after a period in which the power and wealth of the United States were disproportionately important. How the world economy accommodates China and other large, export oriented, rapidly growing countries is of central importance to international economic relations in the twenty-first century.

Vocabulary

BRIC countries	regional trade agreement (RTA)
deep integration	shallow integration
foreign direct investment (FDI)	tariffs
gross domestic product (GDP)	trade-to-GDP ratio
quotas	transaction costs

Study Questions

All problems are assignable in [MyEconLab](#).

1. How can globalization and international economic integration be measured?
2. In what ways is the U.S. economy more integrated with the world today than it was a century ago? In what ways is it less integrated?
3. What does the trade-to-GDP ratio measure? Does a low value indicate that a country is closed to trade with the outside world?
4. Describe the pattern over the last century shown by the trade-to-GDP ratio for leading industrial economies.
5. Trade and capital flows were described and measured in relative terms rather than absolute terms. Explain the difference. Which terms seem more valid—*relative* or *absolute*? Why?
6. In relative terms, international capital flows may not be much greater today than they were a hundred years ago, although they are certainly greater than they were fifty years ago. Qualitatively, however, capital flows are different today. Explain.
7. What are the new issues in international trade and investment? In what sense do they expose national economies to outside influences?
8. Describe the three kinds of evidence that economists use to support the assertion that economies open to the world economy grow faster than economies that are closed.

CHAPTER

14

Comparative Advantage and the Gains from Trade

Learning Objectives

After studying Chapter 3, students will be able to:

- Analyze numerical examples of absolute and comparative advantage.
- Draw a diagram showing gains from trade.
- Define and state the differences between the concepts of absolute advantage, comparative advantage, and competitiveness.
- Discuss the economic and ethical considerations of economic restructuring caused by international trade.

Introduction: The Gains from Trade

This chapter introduces the theory of comparative advantage. A simple model is used to show how nations maximize their material welfare by specializing in goods and services that have the lowest relative costs of production. The improvement in national welfare is known as the **gains from trade**. The concepts of comparative advantage and the gains from trade are two of the oldest and most widely held ideas in all of economics, yet they are often misunderstood and misinterpreted. Therefore, it is worth the effort to develop a clear understanding of both.

Adam Smith and the Attack on Economic Nationalism

The development of modern economic theory is intimately linked to the birth of international economics. In 1776, Adam Smith published *An Inquiry into the Nature and Causes of the Wealth of Nations*, a work that became the first modern statement of economic theory. In the process of laying out the basic ground rules for the efficient allocation of resources, Smith initiated a general attack on **mercantilism**, the system of nationalistic economics that dominated economic thought in the 1700s. Mercantilism stressed exports over imports, primarily as a way to obtain revenues for building armies and national construction projects.

The key mistake in mercantilist thinking was the belief that trade was a **zero sum** activity. In the eighteenth century the term *zero sum* did not exist, but it is a convenient expression for the concept that one nation's gain is another nation's loss. A moment's

reflection should be enough to see the mistake in this belief, at least as it applies to voluntary exchange. When a grocery store sells a gallon of milk or a loaf of bread, both the store and the consumer are better off. If that were not the case, the store would not sell or the consumer would not buy. Voluntary exchanges such as this are positive sum, not zero sum. In this sense, sports metaphors that have a winner and a loser are usually not an apt description of trade relations. Trade is more dance than football, more rock climbing than bicycle racing.

No one in the 1770s thought that they were living in the midst of an industrial revolution, but Smith was observant enough to perceive that many improvements in the standard of living had occurred during his lifetime as a result of increasing specialization in production. When he analyzed specialization, he made one of his most important contributions to economics: the discovery that specialization depends on the size of the market.

A contemporary example may be helpful. If a car company were permitted to sell its cars and trucks only in Michigan, it would have much less revenue and would sell many fewer vehicles. It would hire fewer employees, and each person would be less specialized. As it is, the market is so large (essentially, the world) that car companies can hire engineers who completely specialize in small, even minuscule, parts of a car—door locks, for example. Your door lock engineer will know everything there is to know about the design, production, and assembly of door locks and will be able to put them into cars most efficiently. A firm that was limited to the Michigan market could never afford to hire such specialized skills and could never be as efficient.

One of the keys to Smith's story of wealth creation is access to foreign markets. If no one is willing to import, then every company is limited by the size of the national market. In some cases, that may be large enough (the United States or China), but in most cases, it is not. Small- and medium-sized countries cannot efficiently produce every item they consume. Holland, for example, has always imported a large share of its goods and has depended on access to foreign markets in order to earn export revenues to pay for imports.

Smith was highly critical of trade barriers because they decrease specialization, technological progress, and wealth creation. He also recognized that imports enable a country to obtain goods that it cannot make or cannot make as cheaply, while exports are made for someone else and are useful only if they lead to imports. The modern view of trade shares Smith's dislike of trade barriers for mostly the same reasons. Although international economists recognize that there are limitations to the application of theory, in most cases a majority of economists share a preference for open markets. In Chapters 6 and 7 we will examine trade barriers in greater detail, but at this point we will develop a deeper understanding of the gains from trade by means of a simple algebraic and graphical model.

A Simple Model of Production and Trade

We will begin with one of the simplest models in economics. The conclusion of this analysis is that a policy of free trade maximizes a nation's material

TABLE 14.1 Assumptions of the Simple Ricardian Trade

Model Labor	<ul style="list-style-type: none"> ■ The only input ■ Cannot migrate across borders ■ Is completely mobile between sectors ■ Fully employed
Markets	<ul style="list-style-type: none"> ■ Two outputs ■ Perfect competition ■ No transportation or trade costs
Technology	<ul style="list-style-type: none"> ■ Constant returns to scale ■ No changes in technology or skills

well-being. Later, we will examine some of the cases where real-world conditions do not conform to the assumptions of the model and where the optimality of free trade is questionable.

The basic model is often referred to as a *Ricardian model*, since it first took form in the analysis of David Ricardo. The model begins by assuming that there are only two countries, producing two goods, using one input (labor). The Ricardian model assumes that firms are price takers, or, in other words, markets are competitive, and no firm has market power. The model is static in the sense that it assumes that technology is constant and there are no learning effects of production that might make firms and industries more productive over time. Ricardo also assumed that labor is perfectly mobile between industries but perfectly immobile across national borders. Table 3.1 lists the main assumptions of the model; many of these will be relaxed in later chapters.

Absolute Productivity Advantage and the Gains from Trade

To begin, we define *productivity* in the Ricardian model. Productivity is the amount of output obtained from a unit of input. Since labor is the only input, we can define **labor productivity** as follows:

$$(\text{units of output})/(\text{hours worked})$$

If, for example, two loaves of bread can be produced in one hour, then productivity is as follows:

$$(2 \text{ loaves})/(1 \text{ hour})$$

or two loaves per hour. If four loaves are produced in two hours, then productivity is still as follows:

$$(4 \text{ loaves})/(2 \text{ hours}) = 2 \text{ loaves per hour}$$

Suppose that there are two goods, bread and steel, and two countries, the United States and Canada. Suppose also that each produces according to the productivities shown in Table 3.2.

TABLE 14.2 Output per Hour Worked

	United States	Canada
Bread	2 loaves	3 loaves
Steel	3 tons	1 ton

Canada is more productive than the United States in bread production, but the United States is more productive in steel production.

The values in Table 3.2 show that productivity in the making of bread is greater in Canada than in the United States, and that productivity in steel is greater in the United States. Canada has an **absolute productivity advantage** in bread because it produces more loaves per hour worked (three versus two in the United States). Using the same logic, the United States has an absolute productivity advantage in steel production.

The basis of Adam Smith's support for free trade was the belief that every country would have an absolute advantage in something, and that the source of the advantage did not matter. Whether it was due to special skills in the labor force, climate and soil characteristics of the country, or the temperament of its people, there would be goods that each country could manufacture, grow, or dig out of the ground more efficiently than its trading partner. Consequently, every country could benefit from trade.

In the numerical example outlined in Table 3.2, each loaf of bread costs the United States 1.5 tons of steel. Put another way, the **opportunity cost** of bread is 1.5 tons of steel, since each unit of bread produced requires the economy to move labor out of steel production, forfeiting 1.5 tons of steel that it could have produced instead. This follows from the fact that each hour of labor can produce either 2 loaves of bread or 3 tons of steel. We can write this ratio as the barter price of bread as follows:

$$P_{us}^b = \frac{3 \text{ tons}}{2 \text{ loaves}} = 1.5 \left(\frac{\text{tons}}{\text{loaves}} \right),$$

where b is bread and us is the country. Similarly, we can write the U.S. price of steel as the inverse as follows:

$$P_{us}^s = \frac{2 \text{ loaves}}{3 \text{ tons}} = 0.67 \left(\frac{\text{loaves}}{\text{tons}} \right).$$

You should be able to verify that the Canadian price of bread will be 0.33 (tons/loaf) and that steel will cost 3 (loaves/ton).

If the United States can sell a ton of steel for more than 0.67 loaves of bread, it is better off. Similarly, if Canadians can obtain a ton of steel for fewer than 3 loaves of bread, they are better off. Each country will gain from trade if there is agreement to sell steel for fewer than 3 loaves of bread but more than 0.67 loaves. Anywhere in that range, both Canadians and Americans will benefit. In the end, trade will occur at a price somewhere between these two limits as follows:

$$3.0 \left(\frac{\text{loaves}}{\text{tons}} \right) > P_w^s > 0.67 \left(\frac{\text{loaves}}{\text{tons}} \right),$$

where P_w^s = the world price of steel (the trade price). Without knowing more details about the demand side of the market, it is impossible to say whether the price will settle closer to 3.0 (the Canadian opportunity cost of steel) or 0.67 (the U.S. opportunity cost). The closer the price is to 0.67, the more Canada benefits from trade, and the closer it is to 3.0, the more the United States benefits. Regardless of which country benefits more, as long as the price is between these two limits, both countries benefit from trade.

CASE STUDY

Gains from Trade in Nineteenth-Century Japan

A fundamental result from international economics is that nations gain from trade. We have just shown this in a simple theoretical framework by illustrating how trade enables countries to consume a bundle of goods that is of greater value than what they can produce on their own. The key to this result is that the two trading partners have different productivities, which lead to different prices in autarky.

One question economists have struggled to answer is, “How large are the gains from trade?” Does trade create a relatively small gain, or a relatively large one? The answer is complicated for a couple of reasons. First, there are gains from trade opening that occur immediately and are called *static gains from trade*. But there are also gains that occur over time, called *dynamic gains from trade*, that are difficult to predict since they depend on changes in innovation and productivity. A second reason why it is hard to measure gains from trade is that all countries already trade, so most of what is measured are the potential gains from some additional amount of trade and not the benefits or gains that a country currently has from participating in trade. In our simple model, we went from no trade to some trade, but in the real world, when countries reduce their trade barriers, they go from some trade to some more trade.

Two economists (Berhofen and Brown, *American Economic Review*, 95(1), 2005) tackled this problem in an original way by looking at the case of Japan. Japan’s rulers closed their market to outsiders in 1639 when it felt threatened by Christian missionaries and their Portuguese supporters. From that time on, only the Dutch and the Chinese were permitted to trade with Japan, and each was limited to just a handful of ships per year. In the Dutch case, by the mid-1800s, only one ship per year was allowed to trade, while the Chinese were limited to three or four junks per year. Berhofen and Brown estimate that Japan exported goods worth about 1.2 cents per person and imported even less, around 0.4 cents per person, by the mid-1800s. Essentially, trade was nil and Japan lived in autarky.

As most Americans know from their history books, the United States decided to force open the Japanese market in the early 1850s and sent

Admiral Matthew Perry to accomplish the task. Perry made first contact with Japanese officials in 1853 and signed a limited agreement in 1854. The United States continued to request further opening until a full commercial treaty was signed in 1858 and took effect on July 4, 1859. Following close on the heels of the Americans were the Dutch, Russians, British, and French, and by the mid-1860s, Japan was living under a regime of nearly free trade since its ability to limit imports with tariffs was curtailed by the foreign powers.

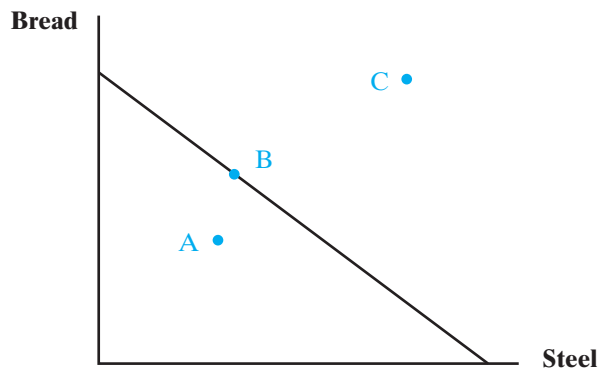
The Japanese case is an excellent one for measuring the static gains from trade. Japan had closed markets before it was forced to open, and after opening it was forced to practice more or less free trade. Our simple theoretical model of trade predicts that Japan should have shifted its domestic production to take advantage of the higher prices offered for its exports, and that its national income should have grown in value since its export goods are worth more and its import goods cost less. Both effects seem to have occurred.

After trade began, Japanese production of silk and tea increased dramatically and these products became major export items. Imported goods included woolen textiles (Japan had no sheep industry) and a variety of manufactured goods, such as weapons, that it did not make itself. National income seems to have grown as well. Berhofen and Brown estimate that an upper bound on the increase was 8–9 percent of GDP. This is not a huge amount, but it is not inconsequential either, and represents only the static gains from trade. Over time, as Japanese producers adjusted to a larger market and as new technologies and products were introduced, additional gains would accrue from increased productivity and innovation.

Comparative Productivity Advantage and the Gains from Trade

At this point, the obvious question to ask is what happens if a country does not have an absolute productivity advantage in anything. It is not hard to imagine an extremely poor, resource-deficient nation with low literacy and scarce capital. What can these countries produce more efficiently than the United States or Germany? Why would a rich country want to trade with them when they are inefficient at everything? The answer is that even if a country lacks a single good in which it has an absolute productivity advantage, it can still benefit from trade. Perhaps even more surprising, high-income countries also benefit from the trade. In other words, the idea that nations benefit from trade has nothing to do with whether a country has an absolute advantage in producing a particular good. In order to see this, first we must develop a few more basic concepts.

FIGURE 14.1 A Production Possibilities Curve for the United States



In a model with only two goods, the production possibilities curve shows the trade-offs.

The Production Possibilities Curve

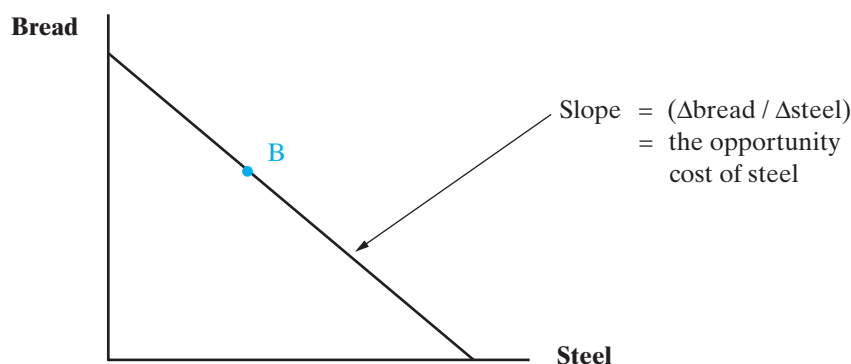
The **production possibilities curve (PPC)** shows the trade-offs a country faces when it chooses its combination of bread and steel output. Figure 3.1 illustrates a hypothetical PPC for the United States. Point B on the PPC is an efficient point of production because it utilizes existing resources to obtain the maximum possible level of output. The assumption of full employment is equivalent to assuming that the United States is operating at a point like B that lies on its PPC. At point A, the economy is inside its production curve and is operating at an inefficient and wasteful level of output because it is not obtaining the maximum possible output from its available inputs. Point C is infeasible because resources do not permit the production of bread and steel in the combination indicated.

The PPC shown in Figure 3.1 is a straight line because it is assumed that the trade-off between bread and steel does not change. This follows from the assumption that labor is homogeneous and that no group of workers is more skilled than another group. The trade-off between bread and steel is another way to refer to the opportunity cost of steel. This follows from the definition of opportunity cost as the best forgone alternative: In order to produce a ton of steel, the United States gives up two-thirds of a loaf of bread. In Figure 3.2, the slope of the PPC is -0.67 , the number of loaves of bread forgone (Δbread) divided by the quantity of steel obtained (Δsteel)—written as follows:

$$\begin{aligned}\text{Slope of the PPC} &= (\Delta\text{bread output})/(\Delta\text{steel output}) \\ &= \text{opportunity cost of steel}\end{aligned}$$

Relative Prices

Suppose that the slope of the PPC is -0.67 , as shown in Figure 3.2. If the United States does not trade, it gives up 0.67 loaves of bread for an additional ton of

FIGURE 14.2 Opportunity Costs and the Slope of the PPC

The slope of the PPC is the opportunity cost of the good on the horizontal axis. This follows from the definition of the slope as the ratio of the vertical change to the horizontal change moving along the PPC.

steel. This trade-off is called the **relative price** of steel or the opportunity cost of steel. The term *relative price* follows from the fact that it is not in monetary units, but rather in units of the other good. If no trade takes place, then the relative price of a good must be equal to its opportunity cost in production.

It is easy to convert the relative price of steel into the relative price of bread: Simply take the inverse of the price of steel. In other words, if 0.67 loaves of bread is the price of 1 ton of steel in the United States, then 1.5 tons of steel is the price of 1 loaf of bread. By the same reasoning, 1.5 tons of steel is the opportunity cost of 1 loaf of bread in the United States when production is at point B or at any other point along the PPC in Figure 3.2.

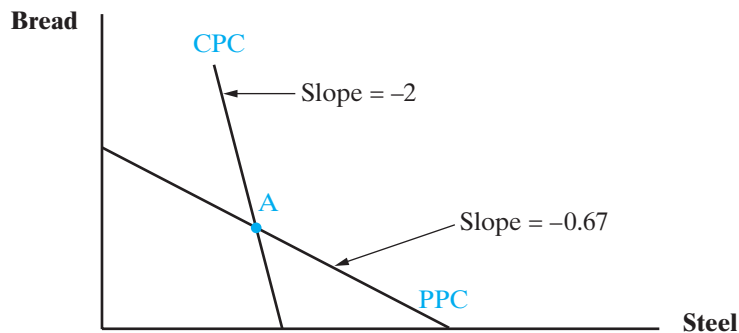
The Consumption Possibilities Curve

The complete absence of trade is called **autarky**, and in this situation both the United States and Canada are limited in their consumption to the goods that they produce at home. Suppose that autarky prevails initially and the opportunity cost of steel in Canada is 3 loaves of bread per ton, and in the United States it is 0.67 loaves per ton (as given in Table 3.1). In this case, both countries can raise their consumption levels if they trade. In particular there will be gains from trade if the price settles somewhere between the opportunity costs in Canada and in the United States. That is, the countries benefit if the following is true:

$$3.0 \text{ (loaves/ton)} > P_w^s > 0.67 \text{ (loaves/ton)}$$

Suppose that the price settles at 2 loaves per ton. In the United States, the pre-trade price was 0.67 loaves per ton. This is illustrated in Figure 3.3, where the PPC for the United States is shown with the production point at A. The trading

FIGURE 14.3 Production and Trade Before Specialization

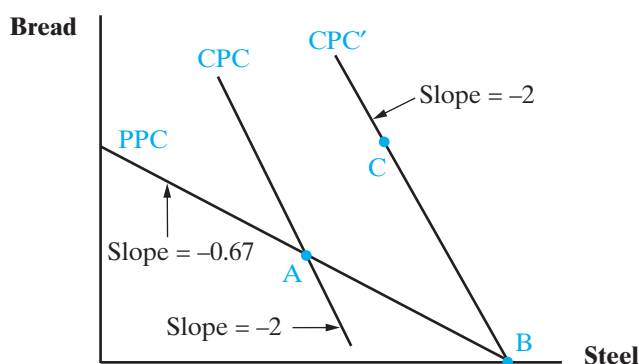


If the United States produces at A and the trade price of steel is 2, then it can trade steel for bread and move its consumption bundle outside its PPC.

possibilities for the United States are illustrated by the **consumption possibilities curve (CPC)**. The slope of the CPC is -2 , which is the relative price of steel, or the rate at which bread and steel can be traded for each other. The CPC passes through point A because this is the combination of steel and bread that is available to trade if the United States produces at A. If the United States chooses to trade, it could move up the CPC, trading each ton of steel for 2 loaves of bread. This is a better trade-off than it gets if it tries to make more bread, since along its PPC each ton brings only two-thirds more loaves of bread. While it is always impossible to produce outside the PPC, in effect, the United States can consume outside it by trading steel for bread.

The Gains from Trade

You should wonder why the United States would choose to make bread at all, since a ton of steel not produced brings in only two-thirds of a loaf of bread. If the United States were to specialize in steel production and trade for bread, it could do much better, since it would get two loaves for each ton. This possibility is shown in Figure 3.4. Here, the pre-trade production point for the United States is at A. This is also its consumption point, since in the absence of trade, consumption must equal production. Point B in Figure 3.4 represents production that is completely specialized in steel. With the opening of trade, production could occur at B, and the United States could trade up along CPC' , which has a slope of -2 , the same as the CPC. If the United States produces at B and moves up CPC' , it can reach a point like C, which is unambiguously superior to the consumption bundle available when production is at point A because it represents more of both bread and steel. Similarly, for any combination of bread and steel that is available along the PPC, or along CPC if the United States produces at A and trades, there is a consumption bundle on CPC' , which represents more of both goods.

FIGURE 14.4 Production to Maximize Income

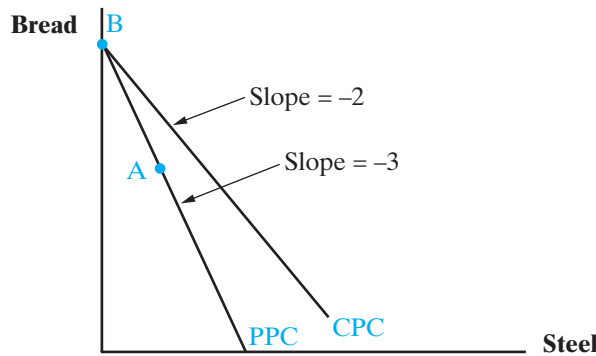
By specializing production at B and trading for bread, the United States obtains the largest possible consumption bundle.

The most important thing to note about production point B is that it maximizes U.S. income. This follows from the fact that it makes available the greatest combinations of bread and steel. To see this, consider that no other point of production puts the United States on a price line that lies farther out from the origin. Every other production point on the United States' PPC lies below the consumption possibilities curve, CPC' , and every CPC with a slope of -2 that passes through the PPC at a point other than B also lies below CPC' . In other words, given the United States' PPC and a relative steel price of 2, the largest bundle of consumption goods is obtained when the United States specializes in steel and trades for its bread.

The United States benefits from trade, but does Canada? Unequivocally, the answer is yes. Consider Figure 3.5 where point A is Canada's pre-trade production point. Along Canada's PPC, the opportunity cost of steel is 3 loaves of bread per ton. After trade, the price settles at 2 loaves per ton. With a trade price of 2, Canada maximizes its income by moving along its PPC to where it is completely specialized in bread production. Then it can trade bread for steel at a trade price that is more favorable than its domestic trade-off of 3 loaves per ton. Canada, too, can consume at a point on CPC that is outside its PPC and above and to the right of its pre-trade equilibrium at point A. Canada, like the United States, is better off because with trade it gets a larger combination of both goods than it can produce for itself.

A numerical example will help clarify the existence of gains from trade. Suppose the relative price of steel is 2 loaves per ton. When the United States increases its steel output by 1 ton, it gives up 0.67 loaves of bread output, but it can trade the steel for 2 loaves, leaving a net gain of 1.33 loaves ($2 - 0.67 = 1.33$). In order to meet U.S. demand for 2 more loaves of bread, Canada must give up 0.67 ton of steel production. It trades the 2 loaves for 1 ton of steel, however, leaving a net gain of 0.33 ton ($1 - 0.67 = 0.33$). Hence, both countries benefit from the trade.

FIGURE 14.5 Canada's Gains from Trade



By specializing production at B and trading for steel, Canada obtains the largest possible consumption bundle.

Domestic Prices and the Trade Price

Now we know that as long as the trade price is between the pre-trade domestic prices in Canada and the United States, both countries can gain from trade. What ensures that the trade price actually settles within this range, $3.0 \text{ (loaves/ton)} > P_w^s > 0.67 \text{ (loaves/ton)}$? What would happen if, for example, P_w^s were equal to 4, or 0.5?

Consider the first case when the trade price is 4 loaves per ton of steel. At $P_w^s = 4$, the trade price of steel is greater than the production cost in each country. Clearly, the United States would want to continue to specialize in steel and trade it for bread. Nothing has changed with regard to the U.S. strategy for maximizing its consumption bundle, or income. The only difference now is that the United States gets 4 units of bread for each unit of steel, instead of 2 as before. In Canada's case, the higher price of steel makes it profitable for Canadian producers to switch to steel production. This follows because the production opportunity cost of steel is 3 loaves of bread, but each ton produced can trade for 4 loaves. By specializing in steel production and trading for bread, Canada maximizes its consumption bundle.

Finally, it should be obvious that both countries are specialized in steel production and that no one is producing bread. There is a bread shortage and a glut of steel. Consequently, bread prices rise, and steel prices fall. This goes on at least until the trade price of steel falls below the opportunity cost of production in Canada, the higher-cost country. Once P_w^s is less than 3, Canadian producers switch back to bread, steel production goes down, bread is up, and trade can resume.

In the second case, where P_w^s is less than 0.67, Canada continues to specialize in bread, and the United States switches. Bread is the surplus good, steel is in short supply, and a similar dynamic causes the price to move in order to ensure that both goods are produced. The equilibrium trade price, then, has to be within

the range we specified earlier, between the opportunity costs in the two countries. In our case, this is between 0.67 and 3.0 loaves per ton.

At the extreme, the trade price could be equal to the opportunity cost in one country; for example, if the trade price of steel is 0.67 loaves per ton, then the United States is indifferent about trading. It cannot be hurt by trading, but it does not gain either, since all the gains go to Canada. Similarly, if the trade price is equal to Canada's opportunity cost, then Canada is indifferent and all the gains accrue to the United States.

Without more information we cannot say much more about the trade price. Will it be close to 0.67 or to 3.0? The answer depends on the strength of demand for each good in both countries, but we have not explicitly included demand in our model, so we cannot say. We do know that if the price is closer to 0.67, then the gains from trade are larger for Canada, and if it is closer to 3.0, the United States benefits more. Nevertheless, both countries gain as long as the price is between the two opportunity costs, and economic forces determine that the price must be in that range.

Absolute and Comparative Productivity Advantage Contrasted

Absolute productivity advantage is defined as having higher labor productivity. We saw that if each country has an absolute productivity advantage in one of the goods, they can both benefit by specializing in that good and trading it for the other good. Note, however, that the gains from trade did not depend in any way on each country having an absolute advantage. In fact, it was the pre-trade opportunity costs of bread and steel in each country that mattered. Opportunity costs were derived from the productivities, but since they are a ratio, vastly different levels of productivity can lead to the same trade-off.

A country has a **comparative productivity advantage** in a good, or simply a comparative advantage, if its opportunity costs of producing a good are lower than those of its trading partners. The concept of comparative advantage is based on the idea that nations maximize their material well-being when they use their resources where they have their highest value. In order to know the highest-valued usage for any resource, we must compare alternative uses. If, by comparison to that of the United States, Canada's opportunity cost of bread is lower, then it should produce more bread and trade for steel.

The distinction between absolute and comparative productivity advantages is one of the most important in economics. It is also one of the least understood, in spite of the fact that it is relatively simple. For example, it is common to read or hear comments about competitiveness that assume that if a country does not have an absolute advantage, it will not be able to sell its products abroad. Our model explains why this logic is erroneous and why even the least productive nations export some goods.

TABLE 14.3 Output per Hour Worked

	Japan	Malaysia
Cars	2	0.5
Steel	2 tons	1 ton

Gains from Trade with No Absolute Advantage

Consider the case shown in Table 3.3. Japan has an absolute advantage in both cars ($2 > 0.5$) and steel ($2 > 1$), yet it can still gain from trade, as can Malaysia, even though it lacks an absolute advantage in either good. If Japan does not trade, it is limited to its own production possibilities, which require it to give up 1 ton of steel for each car it produces. In Malaysia, each car costs 2 tons of steel. Hence, there is scope for a mutually beneficial exchange.

Japan's opportunity cost of steel production is greater than Malaysia's even though it has a higher absolute rate of productivity in steel. Therefore, if it follows its comparative advantage and maximizes its income, it will specialize in cars, the sector where its opportunity cost is lower than Malaysia's. Once trade opens, the world price of cars will be between 1 and 2 tons of steel per car, the opportunity costs of production in Japan and Malaysia, as follows:

$$1\left(\frac{\text{tons}}{\text{car}}\right) < P_w^c < 2\left(\frac{\text{tons}}{\text{car}}\right)$$

Let the price be 1.5 tons of steel per car. If Japan moves to specialize in cars with the opening of trade, it gives up 1 ton of steel for each additional car it produces. With the additional car, it can trade for 1.5 tons of steel, which is a net gain of 0.5 tons over its own production. Similarly, Malaysia gives up 0.5 cars produced for each additional ton of steel it manufactures, but it gains 0.67 cars from each ton of steel traded. Both countries benefit and both countries are able to consume a greater amount of both goods than they could if they relied on their national production alone.

This is a very simplified example of the gains from trade, but it illustrates a fundamental principle. What matters most for the purposes of trade is not a country's absolute advantage, but rather its comparative advantage. This is a central point of international economics: Differences in absolute advantage do not eliminate gains from trade. Furthermore, although both countries gain from trade, it does not imply that their living standards or incomes are equal. Malaysia's income will be less than Japan's because it produces less per hour. In effect, an hour of work in Malaysia returns the equivalent of 1 ton of steel or, through trade, 0.67 cars. Japanese workers produce 2 cars per hour worked, which is equivalent to 3 tons of steel through trade. As a result of higher absolute productivity, incomes in Japan are quite a bit higher, with or without trade.

CASE STUDY

Changing Comparative Advantage in the Republic of Korea, 1960–2007

Few countries began life with a more limited set of possibilities than the Republic of Korea (South Korea). Liberated from its forty years of colonial status (1905–1945) by the defeat of Japan in World War II, Korea was soon wracked by civil war (1950–1953) and divided into two nations. Many observers were pessimistic about the future of noncommunist South Korea. The industrial capacity of the country was mostly located in communist controlled North Korea, and South Korea had little to offer besides the dedication and hard work of its people. Yet, over the following fifty years, few countries have grown faster.

From 1960 to 2010, per capita income in the Republic of Korea grew at the rate of 5.4 percent per year, in real terms (Table 3.4). At this rate, per capita income doubles every thirteen years.

Korea's economic strategy for the first few years after the Korean War was to limit imports and concentrate on producing import substitutes, a common strategy for developing countries in the 1950s. Korea was one of the first to recognize its limitations and to change its policies. In 1960 and 1961, political changes led to a change in economic policies and a more aggressive engagement with the world economy. Korea removed many of its restrictions on imports and began to promote export-oriented industries. Between 1960 and 2010, its trade-to-GDP ratio increased from 15.8 to 102.

Initially, Korea's export efforts were limited to the commodities on hand, mostly minerals, a few agricultural and marine products (for example, seaweed), and very simple consumer goods. Over several decades after 1960, its export industries evolved several times, from simple products requiring few skills and little capital to products that required more skills and greater capital. After its first few years of exploring its comparative advantage, Korea developed competitive sectors in wigs, textiles, shoes, and plywood. With the increase in income came increases in skills and financial capital. This permitted the development of more skill- and capital-intensive industries such as steel, shipbuilding, household appliances, and electronic subassemblies.

TABLE 14.4 Indicators of the Korean Economy

	1960	1980	2000	2010
GDP per capita (\$US, 2000)	1,154	3,358	11,347	16,372
Trade-to-GDP ratio	15.8	72.0	74.3	102.0

Eventually, these were followed by cars, computers, and electronics. By the first decade of the new millennium, Korea was a high-income industrial economy capable of exporting the most technologically advanced products available in several fields. Clearly, its history demonstrates that comparative advantage is not unchangeable, and that it can be a vehicle for raising incomes and promoting development.

An increasing share of Korea's output was sold in world markets. Consequently, production was not limited to the growth in its own domestic market. In addition, its goods had to be competitive in quality and price. Its ability to obtain imports at world prices was also important, but standing behind Korea's competitiveness was its rapid increases in productivity. Without more output per hour of work, incomes could not have risen as fast as they did, and Korea's ability to shift its comparative advantage from low-skill to increasingly higher-skill products could not have gone forward. In turn, productivity increases require a host of complementary changes, ranging from the development of universities and research institutes to organizational changes and the raising of financial capital for investing in new machinery and equipment.

In the process of promoting exports and raising productivity, Korea encountered a number of obstacles including its own bureaucratic inflexibility, problems in marketing to foreign markets that are radically different from Korea's, and a shortage of technical management and industrial expertise. It met and overcame these obstacles, and today Korea is an example of a country that used its comparative advantage to develop its economy. At the same time, it also used the pressure of foreign competition to raise its own productivity and quality standards, which in turn raised the incomes of its citizens. Korea's success was a joint product of efforts by its government, the private sector, and a number of public-private organizations. It is an open question whether each of these played a similar role: Is Korea's success due to the wise guidance of government policies, or did those policies play a secondary (or even negative) role compared to markets and competition?

Comparative Advantage and "Competitiveness"

The rhetoric of "competitiveness" is so common in our public discourse that it is useful to consider its relationship to comparative advantage. In the analysis so far, comparative advantage resulted from productivity differences between nations in autarky. In our simple model of a barter economy, wages, prices, and exchange rates were omitted. Real businesses do not barter steel for bread, however, and they cannot pay their workers by dividing up the firm's output.

In general, by ignoring money wages, money prices, and exchange rates, we assumed that all goods and labor were correctly priced. In other words, we assumed that the prices of outputs and inputs are an accurate indication of their relative scarcity. In this case, there is no difference between a nation's comparative advantage and the ability of its firms to sell goods at prices that are competitive. That is, if all markets correctly value the price of inputs and outputs, then a nation's commercial advantage is determined by its comparative advantage.

Unfortunately, markets sometimes fail to produce optimal outcomes, and at times, outputs and inputs are incorrectly priced. Sometimes, undervaluation or overvaluation of a good stems from inherent difficulties in measuring its true value or in measuring its true cost of production. For example, we usually ignore the costs of air pollution when we measure the costs of driving a car. Other times, undervaluation or overvaluation may result from government policies, as when prices are maintained at an artificially high or low level. In either case, the fact that a market price may not accurately reflect the economic value of an input or an output means that a wedge is driven between commercial or **competitive advantage** and comparative advantage.

It is often (incorrectly) argued that nations should pursue commercial advantages for their firms even if it means a misallocation of resources. In effect, this means that a country follows policies that lower living standards by failing to maximize the value of national output. In terms of Figure 3.4 and Figure 3.5, this is equivalent to asserting that the United States and Canada should each remain at a point like A, where the United States overestimates the value of producing its own bread and Canada overestimates the value of steel. Both countries end up with consumption bundles that are suboptimal from the standpoint of national welfare.

Consider a real-world example. Indonesia tried to develop an aircraft industry in spite of the fact that it lacks a comparative advantage in aircraft production. Nevertheless, through a combination of government policies (some of which paid people to buy the planes!), the price to foreigners was competitive at times. From the perspective of Indonesian national welfare and the optimal use of scarce Indonesian resources, this was a mistake. From the perspective of a business, however, Indonesian policies made it profitable to make airplanes, even though it meant using resources in ways that were suboptimal from the national perspective.

This case illustrates the common mistake of equating nations with business enterprises. Indonesian plane manufacturers care about their subsidies and any other policy that makes them profitable. The national interest, however, is to achieve the most efficient allocation of resources possible within the framework of the nation's laws and values. It is possible to make individual firms highly profitable through subsidies or protection from international competition, while at the same time and through the same policies cause the

nation's overall standard of living to be lower than it would be otherwise. Businesses are not designed to ensure that resources are efficiently allocated at the national level. If they can legally tip the playing field in their direction, they will not hesitate.

Another important distinction between nations and business enterprises is that nations do not compete with each other in any normal sense of the word. Economic relations between the United States and Canada, or any pair of nations, are not equivalent to the commercial competition that exists between companies such as Coke and Pepsi. If Canada grows, the United States does not go out of business or suffer in any identifiable way. In fact, Canadian growth would be a stimulus to U.S. growth and would create spillover benefits for Americans. Cola companies fight over a relatively static market size, but nations can all simultaneously increase their incomes.

Economic Restructuring

Economic restructuring refers to changes in the economy that may require some industries to grow and others to shrink or disappear altogether. For example, the United States has seen a dramatic decrease in the size of its steel industry and, some years later, a rebirth of a new industry based around smaller, more specialized steel mills. Today, the car industry is shrinking, with a long-term prospect that is as yet unknown. In any dynamic economy, some types of economic activity will be growing, and others will be scaling back or dying. In some cases, these changes are a direct consequence of increased openness to foreign competition. For example, the influx of Japanese cars has played a major role in the reorganization and restructuring of the U.S. auto industry.

In our simple Ricardian model, after the opening of trade, the United States was able to maximize its well-being by shifting workers out of bread production and into steel production. Even though this restructuring of the economy improved overall economic welfare, it does not mean that it benefited every individual—a nation's gains from trade may be divided in different ways, and it is usually the case that some individuals benefit while others are hurt by trade. If there are net gains from opening trade (which are measured by an increase in the consumption bundle), then it means that the economic gains of the winners are greater than the economic losses of the losers, and therefore the nation as a whole is better off. Nevertheless, opening an economy to increased foreign competition is rarely painless and usually generates a number of new problems. In the model used in this chapter, it was assumed that workers can effortlessly and without costs move back and forth between industries as one expands and the other shrinks. In reality, this is not an option. While some laid-off workers in a declining industry may quickly find new jobs, many do not. They may not know which companies need workers, or their skills may not match those that are in demand.

The model of comparative advantage does not offer a set of policies for addressing the problems of dislocated workers. Those policies have to come from

another branch of economic analysis, such as labor economics, and from outside economics. It is widely recognized, however, that changes in trade patterns, whether they are due to trade agreements, a unilateral reduction in trade barriers, technological breakthroughs, or any other cause, will result in some dislocation of firms and workers. Most economists continue to support more open trading arrangements, however, because foreign trade increases our choices as consumers, it lowers the costs of inputs for producers, it increases competition and innovation, and it leads to a greater diffusion of technological change. Nevertheless, it is important to keep in mind that the gains from trade do not mean that every worker or every firm benefits.

To a large extent, political assumptions about the way the world works will color the solutions to the problem of worker dislocation offered by economists, political scientists, and other social scientists. For example, believers in less government intervention in the economy would argue that government should not have any policies for handling unemployment caused by the rapid growth of imports. They maintain that unemployment is a self-correcting problem; laid-off workers will look for new jobs and will, if necessary, accept lower wages. Others make a value judgment that this sort of social problem should not be a governmental concern, and that it should be left up to the private economy and individual initiative.

An alternative to the “do nothing” approach is for the government to look for ways to compensate the losers. The proponents of this view justify it on several grounds. First, the nation as a whole benefits from trade, so there are newly added resources to the economy that make compensation possible. Second, many people believe that they have an ethical obligation to assist people hurt by economic change. And third, compensation reduces the incentives to oppose foreign trade.

The practice of offering **trade adjustment assistance (TAA)** is common in many countries, including the United States. Usually these programs take the form of extended unemployment benefits and worker re-training. For example, the U.S. government created a special program of benefits for workers who are hurt by trade with Mexico due to the North American Free Trade Agreement (NAFTA). In 1994, the first year of NAFTA, 17,000 workers qualified for TAA under the NAFTA provision. Generally, in order to qualify for the benefits, workers must demonstrate that they were laid off as a result of imports from Mexico or Canada or because their firm relocated to one of those countries. Needless to say, it is sometimes difficult to establish a direct link between imports and job loss; a poorly managed firm may have been on its way out of business with or without imports.

The important point is that trade creates change, and it may be difficult for some people, industries, or communities to deal with it. When a nation moves along its PPC toward a different mix of industries, there is a period of transition that is painful for some. Economic restructuring does not happen overnight, and although it is desirable for the higher living standard it brings, change and transformation cost time and money.

CASE STUDY**Losing Comparative Advantage**

The case study on Korea shows that comparative advantage is not fixed in time but changes as countries develop their economies. Changing comparative advantage cuts two ways, however, and some production stops being an efficient use of a country's capital and labor. In the Korean case, there are products that it exported early in its development that are no longer cost efficient to make.

Agriculture is an area where many countries experience a declining comparative advantage over time. Some agricultural crops tend to be very labor intensive, and the cost of labor rises as an economy develops. Technology may solve some of the problems of rising wages by reducing the need for labor, but other crops resist an efficient technological solution. In an ideal world, workers in industries that lose their comparative advantage would easily and quickly move to an industry where new opportunities appear.

Comparative advantage in agriculture is not the only concern countries have when thinking about their agricultural sector. Issues of food safety, food independence, and support for rural culture and society are all concerns to one degree or another, more in some countries than others.

One of the objectives of the Doha Round of the World Trade Organization (WTO) is to create an economic environment in which low-cost agricultural producers have access to other countries' markets. The goal is to create greater efficiency in the world economy by locating production where the opportunity costs are lowest, while at the same time creating opportunities for developing countries. If a developing country has a comparative advantage in, say, cotton, but foreign markets are not open, it cannot fully obtain the benefits of its comparative advantage.

Cotton is not a food crop, and its treatment highlights some of the fundamental difficulties involved in persuading countries to drop trade barriers, as well as the fundamental reasons why it is desirable to see barriers fall. According to the International Cotton Advisory Committee, the highest cost producers in the world include Greece, Spain, and the United States, all of which are countries classified as high-income by the World Bank. Lowest-cost producers are in sub-Saharan western Africa (e.g., Burkina Faso, Mali, Benin) and central Asia (e.g., Uzbekistan and Tajikistan).

Cotton is not a major item in world trade, accounting for only about 0.12 percent of total merchandise trade in 2003. Nevertheless, it is important. As many as one hundred million households depend on income earned growing cotton, and several of the low-cost producers depend on their cotton export earnings to buy essential imports such as grain. Table 3.5 compares cotton production, its share of trade, and income per person in a few of the

TABLE 14.5 Low-Cost and High-Cost Cotton Producers

Country	Cotton Exports, 2009 (millions \$)	Percent of Total Exports, 2009	Income per Person, 2009
Low-cost producers			
Western Africa			
Benin	97.6	7.8	772
Burkina Faso	248.7	27.6	509
Mali	73.2	4.1	601
Central Asia			
Tajikistan	71.0	16.0	734
Uzbekistan	259.7	2.4	1,182
High-cost producers			
Greece	402.9	2.0	28,521
United States	3,386.8	0.3	45,793

Sources: Food and Agricultural Organization; World Bank; World Trade Organization.

low-cost and high-cost producers. As shown, low-cost countries produce less but depend more on cotton exports, as their very low levels of income put them close to the edge of survival and they have fewer goods to export. High-cost producers depend much less on their cotton exports and have much higher incomes.

High-cost producers like the United States and Greece depend on a variety of government interventions to keep their cotton producers in business. In Greece, direct and indirect payments, along with tariffs on imports of cotton, are administered through the European Union's Common Agricultural Program. In the United States, the Department of Agriculture administers a number of farm support programs, including payments to farmers, subsidized loans, revenue guarantees, subsidized insurance, marketing and promotion assistance, and others, while the Department of Commerce administers a set of tariffs on foreign cotton entering the U.S. market.

Rich countries that try to keep their high-cost producers in business do more than keep production going where it is less efficient. They also have the potential to harm the living standards of some of the world's poorest countries and to block one of their paths to higher incomes. By using their wealth to subsidize production, high-cost producers increase world supply and limit the ability of low-cost producers to fully exploit their comparative advantage in cotton. In sum, high-income countries find it politically difficult to give up their support for older, less efficient sectors.

Summary

- The single most important determinant of trade patterns is the opportunity cost of producing traded goods. Countries that sacrifice the least amount of alternative production when producing a particular good have the lowest opportunity cost, or a comparative advantage. The idea of comparative advantage has been one of the most enduring concepts of economic thought and has been a central theme in international economic policy since the mid-1800s.
- Nations that produce according to their comparative advantage are maximizing the benefits they receive from trade and, consequently, their national welfare. This is the same as maximizing their gains from trade.
- Comparative advantage is often confused with absolute advantage. The latter refers to the advantage a nation has if its absolute productivity in a particular product is greater than that of its trading partners. It is not necessary to have an absolute advantage in order to have a comparative advantage.
- One common fallacious argument against following comparative advantage is that workers in other countries are paid less than workers at home. This argument neglects the issue of productivity. Developing countries' wages are lower because the value of output from one hour of labor is less. Labor productivity is less because workers are generally less skilled, they have less capital on the job, and they have less capital in the surrounding economy to support their on-the-job productivity.
- Businesspeople look at the issue of trade differently than economists do because they have different objectives in mind. Businesspeople are often concerned about their ability to compete—that is, to sell a particular item in a given market at the lowest price. Their perspective is that of the firm. Economists focus on the efficient use of resources at the national or global level. The perspective is that of all firms taken together.

Vocabulary

absolute productivity advantage	labor productivity
autarky	mercantilism
comparative productivity advantage	opportunity cost
competitive advantage	production possibilities curve (PPC)
consumption possibilities curve (CPC)	relative price
economic restructuring	trade adjustment assistance (TAA)
gains from trade	zero sum

Study Questions

All problems are assignable in [MyEconLab](#).

1. Use the information in the following table on labor productivities in France and Germany to answer questions a through f.

Output per Hour Worked		
	France	Germany
Cheese	2 kilograms	1 kilogram
Cars	0.25	0.5

- a. Which country has an absolute advantage in cheese? In cars?
 - b. What is the relative price of cheese in France if it does not trade? In Germany?
 - c. What is the opportunity cost of cheese in France? In Germany?
 - d. Which country has a comparative advantage in cheese? In cars? Explain your answer.
 - e. What are the upper and lower bounds for the trade price of cheese?
 - f. Draw a hypothetical PPC for France and label its slope. Suppose that France follows its comparative advantage in deciding where to produce on its PPC. Label its production point. If the trade price of cars is 5 kilograms of cheese per car, draw a trade line (CPC) showing how France can gain from trade.
2. Suppose that the table in Study Question 1 looks as follows. Use the information to answer questions a through f.

Output per Hour Worked		
	France	Germany
Cheese	1 kilogram	2 kilograms
Cars	0.25 car	2 cars

- a. Which country has an absolute advantage in cheese? In cars?
- b. What is the relative price of cheese in France if it does not trade? In Germany?
- c. What is the opportunity cost of cheese in France? In Germany?
- d. Which country has a comparative advantage in cheese? In cars? Explain your answer.

- e. What are the upper and lower bounds for the trade price of cheese?
 - f. Draw a hypothetical PPC for France and label its slope. Suppose that France follows its comparative advantage in deciding where to produce on its PPC. Label its production point. If the trade price of cars is 5 kilograms of cheese per car, draw a trade line (CPC) showing how France can gain from trade.
3. Explain how a nation can gain from trade even though as a result not everyone is better off. Is this a contradiction?
 4. Economic nationalists in developed countries worry that international trade is destroying the national economy. A common complaint is that trade agreements open the economy to increased trade with countries where workers are paid a fraction of what they earn at home. Explain the faulty logic of this argument.
 5. Many people believe that the goal of international trade should be to create jobs. Consequently, when they see workers laid off due to a firm's inability to compete against cheaper and better imports, they assume that trade must be bad for the economy. Is this assumption correct? Why, or why not?
 6. Suppose that Germany decides to become self-sufficient in bananas and even to export them. In order to accomplish these goals, large tax incentives are granted to companies that will invest in banana production. Soon, the German industry is competitive and able to sell bananas at the lowest price anywhere. Does Germany have a comparative advantage? Why, or why not? What are the consequences for the overall economy?

CHAPTER

15

Exchange Rates and Exchange Rate Systems

Learning Objectives

After studying Chapter 10, students will be able to:

- List the reasons for holding foreign exchange and the main institutions in the foreign exchange market.
- Differentiate short-run, medium-run, and long-run forces that help determine the value of a currency.
- Diagram the effects on the home currency of a change in supply or demand for foreign currency.
- Calculate a currency's forward premium or discount based on interest rate differentials.
- Use a simple equation to demonstrate the effect on the real exchange rate of higher inflation at home.

Introduction: Fixed, Flexible, or In-Between?

Every country must choose an exchange rate system to determine how prices in the home country currency are converted into prices in another country's currency. Some countries peg their exchange rate at a fixed level while others let market forces determine the value of their currency. Both approaches have advantages and disadvantages. The choice of an exchange rate system varies along the continuum from completely fixed with no variation to completely flexible with variation determined by supply and demand for the country's currency on a minute-by-minute basis. Between these two extremes are several other exchange rate systems with semi-fixed or semi-flexible rates.

Each exchange rate system requires that governments and central banks have credible policies to support the selected system as trade, capital flows, and other pressures from the world economy push exchange rates up and down. In this chapter we define the actors in currency markets, analyze the basic mechanisms that determine the value of a country's currency, and discuss the considerations that countries should make in selecting their exchange rate system. Each of these elements is an important determinant of a country's exchange rate system and the value of its currency.

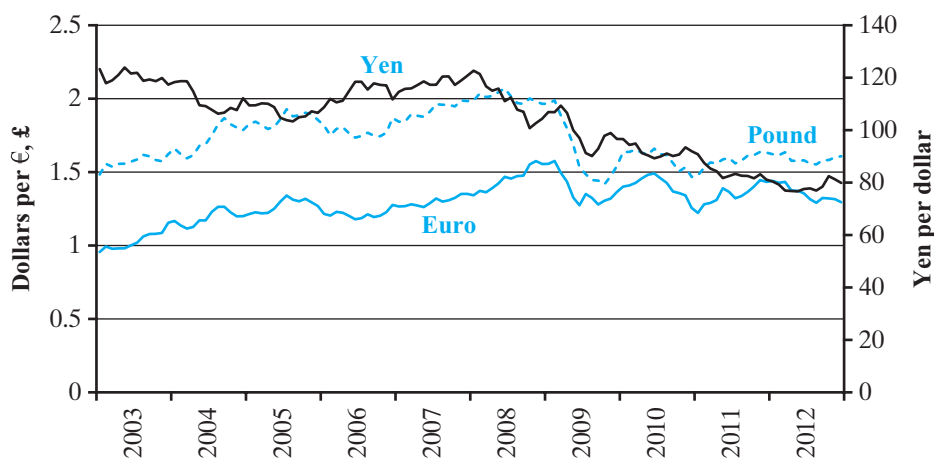
Exchange Rates and Currency Trading

The **exchange rate** is the price of one currency stated in terms of a second currency. An exchange rate can be given in one of two ways, either as units of domestic currency per unit of foreign currency or vice versa. For example, we might give the U.S.-Mexico exchange rate as dollars per peso (0.10 dollars) or pesos per dollar (10 pesos). The custom varies with the currency. For example, the U.S. dollar-British pound exchange rate is usually quoted in terms of dollars per pound, but the U.S. dollar-Mexican peso exchange rate is usually pesos per dollar. In this chapter and the rest of the book, the exchange rate is always given as the number of units of domestic currency per unit of foreign currency. For the United States, this means it is dollars per peso and dollars per pound.

Exchange rates are reported in every newspaper with a business section and on numerous Web sites. Figure 10.1 shows several years of U.S. dollar values for three of the most frequently traded currencies: the European Union's euro, the Japanese yen, and the British pound. The rates are taken from the U.S. Federal Reserve Bank's Web site, and are the interbank rates that one bank charges another when buying large amounts of currency. Tourists and individuals purchasing relatively small amounts would have paid more.

All three are flexible exchange rates, meaning that they are not fixed over time. Figure 10.1 shows a steady appreciation of the euro and the pound against the dollar, from the beginning of the series until shortly after the financial crisis and recession of 2007–2009. At about the same time, the Japanese yen strengthened appreciably (fewer yen per dollar). Each of the three series underwent a change of

FIGURE 15.1 Dollar Exchange Rates for Three Main Currencies, 2002–2012



Euro and pound exchange rates are on the left scale; yen are on the right. Floating exchange rates can vary significantly, in both the short and the long run.

Source: Board of Governors of the Federal Reserve System.

between one-third and one-half between their maximum and minimum values. Considering that the changes in the real economies of each country were not nearly as large, Figure 10.1 gives an idea of the variability of exchange rates.

At the end of the series, the Japanese yen is at a level equal to approximately 80 yen per dollar, or 1.25 cents per yen. The euro is \$1.30 and the pound is \$1.60. It is tempting to conclude that the yen is weak and the pound and euro are strong given that the yen is barely over 1 cent and other two are well over 1 dollar. This conclusion would be mistaken, however, because currencies are scales, just like Fahrenheit or centigrade, or miles and kilometers. No matter how many of one it takes to equal another, we cannot conclude that the scale implies strength or weakness. By way of illustration, consider that the yen has appreciated in value against the dollar and the pound since mid-2007 and against the euro since mid-2008. In reality, the yen has been the strongest currency over the last several years.

Reasons for Holding Foreign Currencies

Economists identify three reasons for holding foreign currency. The first is for trade and investment purposes. Traders (importers and exporters) and investors routinely transact in foreign currencies, either receiving or making payments in another country's money. Tourists are included in this category because they hold foreign exchange in order to buy foreign goods and services.

The second reason for holding foreign exchange is to take advantage of interest rate differentials, or **interest rate arbitrage**. Arbitrage conveys the idea of buying something where it is relatively cheap and selling it where it is relatively expensive. Interest rate arbitrage is similar in that arbitrageurs borrow money where interest rates are relatively low and lend it where rates are relatively high. By moving financial capital in this way, interest rate arbitrage keeps interest rates from diverging too far, and also constitutes one of the primary linkages between national economies. Over the last several years interest arbitrageurs have played a major role in keeping the Japanese yen strong by borrowing in Japan where interest rates are very low and lending where they are high. Various other factors, such as perceptions of risk, are important, but in general, interest rate arbitrage is a powerful force in the world economy and tends to be one of the main reasons for holding foreign currency.

The third reason for holding foreign exchange is to speculate. Speculators are businesses that buy or sell a currency because they expect its price to rise or fall. They have no need for foreign exchange to buy goods or services or financial assets; rather, they hope to realize profits or avoid losses through correctly anticipating changes in a currency's market value. Speculators are often reviled in the popular press, but in fact they help to bring currencies into equilibrium after they have become over- or undervalued. If speculators view a currency as overvalued, they will sell it and drive down its value. If they guess wrong, however, they can lose a lot of money. For this reason, some economists have argued that speculation either serves the useful function of bringing currency values into proper alignment, or its practitioners lose money and go out of business. Not everyone agrees with this view, however, and some economists think that speculation against a currency can

be destabilizing in the sense that it does not always push an exchange rate to its equilibrium value, but instead sometimes leads to a grossly over- or undervalued currency, which is a major problem for the country involved.

Institutions

There are four main participants in foreign currency markets: retail customers, commercial banks, foreign exchange brokers, and central banks. Of these four, commercial banks are the most important. Retail customers include firms and individuals that hold foreign exchange for any of the three reasons given in the previous section—to engage in purchases, to adjust their portfolios, or to profit from expected future currency movements. In most cases, they buy and sell through a commercial bank. Commercial banks in many parts of the world hold inventories of foreign currencies as part of the services offered to customers. Not all banks provide this service, but those that do usually have a relationship with several foreign banks where they hold their balances of foreign currencies. When a surplus accumulates, or a shortage develops, the banks trade with each other to adjust their holdings.

In the United States, foreign exchange brokers also play an important role. It is not very common for U.S. banks to trade currency with foreign banks. Instead, U.S. banks tend to go through foreign exchange brokers, who act as middlemen between buyers and sellers that do not usually hold foreign exchange. Brokers can also serve as agents for central banks. The market, then, works as follows. An individual or firm that needs foreign exchange calls its bank. The bank quotes a price at which it will sell the currency. The price is based on one of two possible sources of supply: The bank may have an account with another bank in the country where the currency is used, or it may call a foreign exchange broker. The broker keeps track of buyers and sellers of currencies and acts as a deal maker by bringing together a seller and a bank that is buying for its customer.

In most cases, currency trades take the form of credits and debits to a firm's bank accounts. For example, a local U.S. importer that must make payment in yen can call and tell its bank to transfer yen to the Japanese bank of the firm that supplies the importer with goods. The importer will have a debit to its local bank account that is equivalent to the cost of the yen. If the U.S. bank has a branch or correspondence bank in Japan, it can electronically notify the branch to debit the yen from the account of the U.S. bank and credit it to the Japanese bank of the supplier. If the U.S. bank goes through a currency trader instead of dealing directly with a Japanese bank, it first buys yen that are in an account with a Japanese bank. Next, it requests that some or all of its yen assets be transferred to the bank of the Japanese supplier of the U.S. importer.

Exchange Rate Risk

Firms that do business in more than one country are subject to **exchange rate risks**. These risks stem from the fact that currencies are constantly changing in value and, as a result, expected future payments that will be made or received in a foreign currency will be a different domestic currency amount from when the contract was signed.

Suppose, for example, that a U.S. semiconductor manufacturer signs a contract to send a British computer manufacturer a shipment of microprocessors in six months. If the U.S. manufacturer agrees on a price in British pounds, it must know the value of the pound six months from now in order to know the dollar equivalent of its future revenue. If the U.S. manufacturer specifies that the microprocessors be paid for in dollars, then it shifts the exchange rate risk to the British firm. The U.S. company knows the exact dollar amount it will receive in six months, but the British firm is uncertain of the price of the dollar and therefore the pound price of microprocessors.

Financial markets recognized this problem long ago and, in the nineteenth century, they created mechanisms for dealing with it. The mechanisms are the forward exchange rate and the forward market. The **forward exchange rate** is the price of a currency that will be delivered in the future; the **forward market** refers to the market in which the buying and selling of currencies for future delivery takes place. Forward markets for currencies are an everyday tool for international traders, investors, and speculators because they are a way to eliminate the exchange rate risk associated with future payments and receipts. Forward foreign exchange markets allow an exporter or importer to sign a currency contract on the day they sign an agreement to ship or receive goods. The currency contract guarantees a set price for the foreign currency, usually 30, 90, or 180 days into the future. By contrast, the market for buying and selling in the present is called a **spot market**. The prices of foreign currencies quoted in Figure 10.1 are “spot prices.”

Suppose the U.S. semiconductor manufacturer signs a contract to deliver the microprocessors to the British firm in six months. Suppose also that the price is stated in British pounds. The manufacturer knows precisely how many pounds it will earn six months from now, but it does not know whether the pound will rise or fall in value, so it does not know what it will earn in dollar terms. The solution is to sign a forward contract to sell British pounds six months from now in exchange for U.S. dollars at a price agreed upon today. Using the forward market, the U.S. manufacturer avoids the risk that comes from exchange rate fluctuations.

Forward markets are important to financial investors and speculators as well as exporters and importers. For example, bondholders and other interest rate arbitrageurs often use forward markets to protect themselves against the foreign exchange risk incurred while holding foreign bonds and other financial assets. This is called **hedging** and it is accomplished by buying a forward contract to sell foreign currency at the same time that the bond or other interest-earning asset matures. When interest rate arbitrageurs use the forward market to insure against exchange rate risk, it is called **covered interest arbitrage**.

The Supply and Demand for Foreign Exchange

The value of one nation’s money, like most things, can be analyzed by looking at its supply and demand. Under a system of flexible, or floating, exchange rates, an increase in the demand for the dollar will raise its price (cause an **appreciation** in

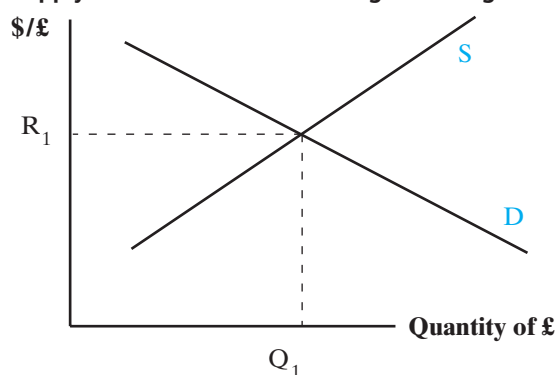
its value), while an increase in its supply will lower its price (cause a **depreciation**). Under a fixed exchange rate system, the value of the dollar is held constant through the actions of the central bank that counteract the market forces of supply and demand. Consequently, supply and demand analysis is a useful tool for understanding the pressures on a currency regardless of the type of exchange rate system adopted. For this reason, we begin with the assumption that exchange rates are completely flexible. After examining the usefulness of supply and demand analysis, we will turn to alternative systems, including gold standards and other variations on fixed exchange rates.

Supply and Demand with Flexible Exchange Rates

Figure 10.2 shows the supply and demand for British pounds in the United States. The demand curve is a normal, downward sloping curve, indicating that as the pound depreciates relative to the dollar, the quantity of pounds demanded by Americans increases. Note also that we are measuring the price of the pound—the exchange rate—on the vertical axis. Since it is dollars per pound ($\$/\text{£}$), it is the price of a pound in terms of dollars, and an increase in the exchange rate (R) is a decline in the value of the dollar. Movements up the vertical axis represent an increase in the price of the pound, which is equivalent to a fall in the price of the dollar. Similarly, movements down the vertical axis represent a decrease in the price of the pound.

British goods are less expensive for Americans when the pound is cheaper and the dollar is stronger. Hence, at depreciated values for the pound, Americans will switch from U.S. or third-party suppliers of goods and services to British suppliers. However, before they can purchase goods made in Britain, first they must exchange dollars for British pounds. Consequently, the increased demand for British goods is simultaneously an increase in the quantity of British pounds demanded.

FIGURE 15.2 Supply and Demand in the Foreign Exchange Market



The intersection of the supply of British pounds to the U.S. market and the U.S. demand for British pounds determines the quantity of pounds available in the United States (Q_1) and their dollar price (exchange rate R_1).

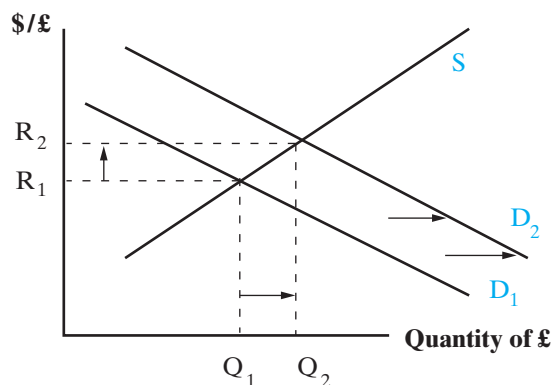
The supply curve in Figure 10.2 slopes up because British firms and consumers are willing to buy a greater quantity of American goods as the dollar becomes cheaper. That is, they receive more dollars per pound. However, before British customers can buy American goods, first they must convert pounds into dollars, so the increase in the quantity of American goods demanded is simultaneously an increase in the quantity of foreign currency supplied to the United States. The intersection of supply and demand determines the market exchange rate and the quantity of pounds supplied to the United States. At exchange rate R_1 , the demand and supply of British pounds to the United States is Q_1 .

Exchange Rates in the Long Run

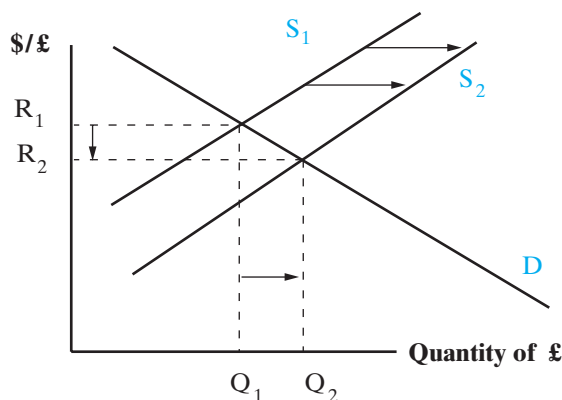
We have determined that the supply curve slopes up to the right and the demand curve slopes down. The next step in supply and demand analysis is to consider the factors that determine the intersection of supply and demand and the actual exchange rate. We will continue to assume that the exchange rate is completely flexible. Later in the chapter we look at exchange rates that are fixed, and at intermediate rates between fixed and flexible.

In Figure 10.3, an increase in the U.S. demand for the pound (rightward shift of the demand curve) causes a rise in the exchange rate, an appreciation in the pound, and a depreciation in the dollar. Conversely, a fall in demand would shift the demand curve left and lead to a falling pound and a rising dollar. On the supply side, an increase in the supply of pounds to the U.S. market (supply curve shifts right) is illustrated in Figure 10.4, where a new intersection for supply and demand occurs at a lower exchange rate and an appreciated dollar. A decrease in the supply of pounds shifts the curve leftward, causing the exchange rate to rise and the dollar to depreciate.

FIGURE 15.3 An Increase in Demand for British Pounds



An increase in the U.S. demand for British pounds (rightward shift of the curve) causes the dollar to depreciate.

FIGURE 15.4 An Increase in the Supply of British Pounds

An increase in the supply of British pounds to the U.S. market (rightward shift of the curve) causes the dollar to appreciate.

The causal factors behind the shifts in the supply and demand are easier to conceptualize if we divide the determinants of exchange rates into three periods: long run, medium run, and short run. This seems to be accurate empirically, as not all the factors that determine an exchange rate show up instantaneously. In fact, some causal factors take a very long time—a decade or more—to exert their full influence, and in the meantime, a number of short-run or medium-run factors may push in a completely opposite direction.

Looking at the long run first, **purchasing power parity** states that the equilibrium value of an exchange rate is at the level that allows a given amount of money to buy the same quantity of goods abroad that it will buy at home. By this criterion, the equilibrium exchange rate is the point where the dollar buys pounds at a rate that keeps its purchasing power over goods and services constant. That is, \$100 buys the right amount of pounds to enable the purchase of the same basket of goods and services in Britain that \$100 buys in the United States. Table 10.1 illustrates this idea.

TABLE 15.1 A Hypothetical Example of the Exchange Rate in the Long Run

	Cost of the Same Basket of Goods in Each Country
Price in dollars	\$1,000
Price in pounds	£500
Long-run equilibrium exchange rate	$(\$1,000/£500) = \$2/£$
Purchasing power parity states that dollars will tend to exchange for pounds at a rate that maintains a constant purchasing power of a given quantity of currency.	

In Table 10.1, a hypothetical basket of goods costs \$1,000 or £500, depending on the country where it is purchased. Accordingly, the long-run tendency is for the exchange rate to move to \$2 per pound. If it is above that, the pound is overvalued and the dollar is undervalued. An overvalued pound buys more in the United States than in Britain since it would be possible to convert £500 to more than \$1,000 and buy a larger basket of goods than can be bought in Britain. Exchange rates less than \$2 would imply the opposite—the pound is undervalued and the dollar overvalued.

It should be stressed that this is an underlying tendency and not a description of actual exchange rates at any point in time. Over the long run, purchasing power parity exerts influence over exchange rates, but in the short to medium run, there are significant deviations from this pattern. If you have traveled outside your home country, you are probably aware of cases where your domestic currency buys you so much foreign currency that your standard of living is higher when you travel. You might be able to stay in a better class of hotel, eat in better restaurants, and shop for items that you cannot afford at home; or you may be familiar with the opposite scenario, where your standard of living declines because you get so little foreign currency in exchange for your domestic currency that everything seems inordinately expensive.

Purchasing power parity influences currency values indirectly. When a currency is overvalued or undervalued, it creates profit-making opportunities for merchants that can move goods across international borders. Suppose, for example, that the dollar is overvalued and that instead of \$2 per pound, the exchange rate is \$1.75 per pound. Prices are assumed to be the same as those shown in Table 10.1. In this case, \$1,000 buys £571.43 ($1,000/1.75$). If merchants take the £571.43 and buy British goods and then ship the goods to the United States, they can earn more than \$1,000. (They earn \$1,142.86 since goods prices are 2 to 1.) In the long run, the demand for British pounds increases and, as shown in Figure 10.3, the exchange rate rises. The process will continue until the exchange rate hits \$2 per pound and there are no more profit-making opportunities from shipping goods from Britain to the United States.

The process just described is reinforced by the flow of goods from Britain to the United States. The supply of goods shrinks in Britain, leading to rising prices there. In the United States, supply rises and, under normal competitive conditions, prices will fall. These effects will take a while to exert themselves, but they are another factor reinforcing purchasing power parity. In this case, however, prices are moving in the direction that equalizes the purchasing power of the two currencies instead of equalization through exchange rate movement as in the previous example. In theory, it does not matter which changes—prices or exchange rates—but given that prices in many countries tend not to fall easily, while exchange rates are relatively easily moved, most of the equalization probably occurs through exchange rate movements.

The story of goods arbitrage—buying where the goods are cheaper and selling where they are more expensive—which stands behind purchasing power parity, obviously has a few unrealistic assumptions. In particular, it requires that goods

flow costlessly across international borders and that all goods and services can be traded. In reality, there are transportation costs involved with moving goods. This means that our merchant who buys £571.43 of goods in Britain and sells them for \$1,142.86 in the United States loses some of his or her \$142.86 profit to shipping, insurance, and other transaction costs. In addition, he or she pays a fee to a bank or a currency broker when buying the needed pounds.

Nor is this the only obstacle standing in the way of profits. Few nations have eliminated all their barriers to the entry of foreign goods and services. The merchant may face a tariff, import license fees, inspection fees, or some other barrier at the border that adds to his or her cost. In the limit, imports of the goods in question may be prohibited and goods arbitrage may be impossible at any price differential. In addition, some goods and many services are not traded. For example, restaurant meals, haircuts, landscape maintenance, and a host of other services that must be consumed on the spot are rarely, if ever, traded.

Once the assumptions of purchasing power parity are examined, it is not surprising that it exerts its influence over exchange rates only in the long run. If there are significant profit-making opportunities through goods arbitrage, then in spite of today's obstacles, entrepreneurs will work to create the conditions that will allow them to take advantage of the price differentials across markets. They will look for ways to lower transport costs, to minimize the costs of compliance with import rules and regulations, and to change the rules where it is feasible. All of these steps take time, but in spite of the real obstacles to its operation, purchasing power parity remains a significant long-run force in the determination of exchange rates.

Exchange Rates in the Medium Run and Short Run

While purchasing power parity is working slowly in the background, other forces have more immediate impacts on the position of the supply and demand curves for foreign exchange. We turn first to the forces that are correlated with the business cycle, the natural but irregular rhythms of expansion and recession that every country undergoes. Given that the time period from the peak of one expansion to the next is usually several years in duration, the forces that are tied to the business cycle can be considered medium run. That is, they are pressures on an exchange rate that may last for several years, but almost always less than a decade and usually less than five to seven years.

The most important medium-run force is the strength of a country's economic growth. Rapid growth implies rising incomes and increased consumption. When consumers feel secure in their jobs and at the same time experience a rapid growth in their incomes, they spend more, some of which will be on imports and travel abroad. As a result, rapid economic growth at home is translated into increased imports and an outward shift in the demand for foreign currency, as shown in Figure 10.3. Holding constant a host of short-run forces that may be in play at the same time, the effect of rapid economic growth at home is a depreciating currency.

The effect of growth is symmetrical, both with respect to slower growth at home, and with respect to the rate of economic growth abroad. Slower growth, such as a recession during which output declines (negative economic growth), raises consumer uncertainty about jobs and reduces many people's incomes. For the economy as a whole, as consumption expenditures fall, expenditures on imports decline as well, and the demand for foreign exchange falls. A leftward shift of the demand curve reduces the exchange rate and appreciates the currency. In other words, just as more rapid economic growth can cause a depreciation in a country's currency, slower growth sets forces in motion that lead to an appreciation.

Growth abroad does not have a direct effect on the home country's demand for foreign exchange (although it may have an indirect effect through its stimulation of the home economy), but it will directly affect the supply curve. More rapid foreign growth leads to more exports from the home country, and slower foreign growth results in fewer exports. More exports to foreigners increase the supply of foreign currency and shift the supply curve rightward, as shown in Figure 10.4. Fewer exports have the opposite effect. You should practice drawing the effects of changes in the rates of home and foreign economic growth on the supply and demand curves for foreign exchange.

Turning from the medium run of the business cycle to short-run periods of a year or less, a number of forces are constantly at work shaping currency values. The foremost short-run force is the flow of financial capital. The effects of financial flows range from minor and subtle to dramatic and, at times, catastrophic. They are as capable of creating slight day-to-day variations in the value of a currency as they are of creating complete financial chaos and bringing down governments. The degree of volatility in financial flows varies greatly and is highly responsive to governmental policies and conditions in the world economy. The impact on exchange rates of large-scale, short-run movements in financial capital has become one of the most serious issues in international economics.

Two variables in particular are responsible for a large share of short-run capital flows: interest rates and expectations about future exchange rates. These two forces often influence each other and are capable of creating unpredictable interactions, as when a change in interest rates reshapes investor confidence or catalyzes speculative actions in currency markets.

The role of interest rates in the short-run determination of exchange rates is crucial. The interest rate–exchange rate relationship is summed up in the **interest parity** condition, which states that the difference between any pair of countries' interest rates is approximately equal to the expected change in the exchange rate. The appendix at the end of this chapter develops the algebra of this relationship, but the intuition is not difficult to grasp. Suppose an investor has a choice between investing at home and earning interest i , or investing abroad and earning interest rate i^* . If foreign interest rates are higher than domestic ones, it may seem advantageous to invest abroad, but this is not necessarily the case. The best choice is also determined by exchange rate movements during the investment period. If investors want to convert their future earnings back into their home currency,

then exchange rate movements must be taken into account during the investment period. To protect against unanticipated losses due to currency fluctuations, cross border investors can sign a forward contract to sell the foreign exchange from their future earnings. This is known as covered interest arbitrage and is a common way to take advantage of interest differentials while guarding against the risk of exchange rate losses.

A simple example will help clarify. Suppose a U.S.-based investor has a choice between a one-year certificate of deposit (CD) issued by a U.S. bank or a German bank. For the sake of simplicity, assume that the CDs are similar with respect to risk, transaction costs, and other characteristics. The U.S. investment is denominated in dollars and pays 3 percent (i) while the German investment is in euros and pays 2 percent (i^*). In one year, \$1,000 invested in the United States will pay $\$1,000 \times (1 + 0.03)$, or \$1,030, while the return on the German CD depends on the fixed interest rate and the exchange rate a year from now. If the dollar-euro spot rate is 1.2 today, then the investor can use the \$1,000 to buy $(1,000/1.2)$, or €833.33, which can be invested at 2 percent in Germany. In one year, the investor will have $833.33 \times (1 + 0.02)$, or €850. If the exchange rate is 1.3 a year from now, then \$1,000 converted to euros today and invested in Germany will pay 850×1.3 , or \$1,105. That is, the investor earns $(1.3/R) \times (1 + 0.02)$ in one year, where R is today's spot rate of exchange.

The problem for the investor is that he or she cannot know what the exchange rate will be one year from now. Our example fudged this point by assuming that the rate was 1.3 dollars per euro in one year's time, but in fact we cannot know what the spot exchange rate will be in a year. Given this uncertainty, investors turn to the forward market where they can sign a contract guaranteeing them a fixed amount of dollars for the euros they will have in one year when the CD matures.

Let F stand for the forward exchange rate and R for the spot rate. The difference between the two is the expected appreciation or depreciation. If $F > R$, then the dollar is expected to depreciate, and is said to be selling at a discount. If $F < R$, then the dollar is expected to appreciate and is selling at a premium. Given information about F and R , our investor is prepared to select between the dollar and euro CDs.

In our previous example, R is 1.2 and F is 1.3, implying that the dollar is at a discount in the forward market and people expect it to depreciate over the next year. The choices are as follows. An investor with \$1,000 can earn $\$1,000 \times (1.03)$, or \$1,030 in the United States, or he or she can earn $(1.3/1.2) \times (1.02) \times \$1,000$, which is \$1,105 if he or she invests in Germany. Clearly the German investment is better and will attract capital. Money flowing into German CDs will push down German interest rates (i^* falls) and increase the spot price of the euro (R rises). Both changes reduce earnings on the German CD until, in the end, we reach the interest parity condition

$$i - i^* \approx (F - R)/R$$

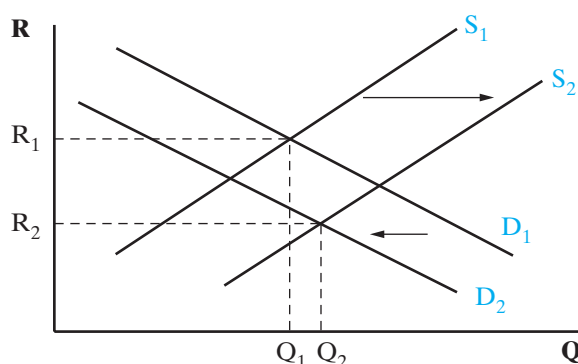
which says that interest rate differences are approximately equal to the expected change in the exchange rate.

The utility of the interest parity condition is that it brings together capital flows, domestic interest rate policy, and exchange rate expectations. Suppose, for example, that domestic interest rates are above foreign rates, so that $i > i^*$. In that case, investors expect a discount in the forward market, so that $F > R$. If the expected depreciation in the domestic currency is not sufficient to compensate for higher interest rates at home, then capital flows into the home country and increases demand for domestic currency, pushing down domestic interest rates until the difference between i and i^* is approximately equal to the percentage difference between the forward and spot exchange rates.

Consider another example. Suppose that home interest rates are less than foreign rates ($i < i^*$) and that forward rates are less than spot rates ($F < R$) by an appropriate amount so that the interest parity condition holds. Beginning at this point, home policymakers decide for some reason to raise their interest rates to the same level as foreign rates: $i = i^*$. Now, investors in both home and foreign markets will invest more at home because they earn the same rate of interest, and they expect domestic currency to appreciate in value (since $F < R$). Figure 10.5 illustrates these shifts. Note that both the demand curve for foreign currency and the supply curve of foreign currency shift, with demand moving in and supply moving out. Taken together, both shifts reinforce a downward movement in the spot rate. As R falls, domestic currency appreciates and the gap between F and R closes. If $i = i^*$, the process ends when $F = R$.

In addition to their impact on the forward-spot rate differential, expectations play a crucial role in the determination of exchange rates in another way. A sudden change in the expected future value of an exchange rate can have a dramatic and often self-fulfilling impact on a country's currency. For example, if investors suddenly come to believe that a currency must depreciate more than they had anticipated, it lowers the expected value of assets denominated in that currency.

FIGURE 15.5 The Effects of an Increase in Home's Interest Rate



An increase in domestic interest rates causes a decrease in demand and an increase in supply of the foreign currency. Both effects cause an appreciation in the exchange rate from R_1 to R_2 .

This can create a sudden exodus of financial capital and put enormous pressure on the country's supply of foreign exchange reserves. To a significant extent, episodes of capital flight can be self-fulfilling in their expectations about an exchange rate. If investors expect depreciation, they try to convert their assets to another currency. This raises the demand for foreign exchange and depresses the supply, fulfilling the expectation of a depreciation.

There are numerous potential causes of this type of volatility in financial capital flows and exchange rate shifts. It also seems likely that technological changes in telecommunications have altered the sensitivity of markets toward changes in expectations, although this is yet to be established definitively. Nevertheless, it is certain that a frequent cause of sudden shifts in expectations is the realization that a particular government is practicing economic policies that are internally inconsistent and unsustainable. We will examine this in more detail in Chapters 11 and 12, but it is relatively easy to get a sense of the meaning of inconsistent policies. An example is policies that are designed to stimulate the economy strongly (more growth → more imports → more demand for foreign exchange) when the supply of foreign exchange is severely limited (not enough exports, very low interest rates).

CASE STUDY

The Largest Market in the World

In 2010, the world's foreign exchange markets traded an estimated \$3,981 billion worth of currency per day. Another way to look at this is that every 3.6 days, currency trades equaled the value of U.S. annual GDP. These estimates come from a survey of fifty-three central banks conducted every three years by the Bank for International Settlements (BIS), a "central bank for central banks." The BIS survey is the *Triennial Central Bank Survey of Foreign Exchange and Derivatives Market Activity in April 2010*, and is available from the BIS at <http://www.bis.org>.

Between 1992 and 2010, the volume of exchange rate transactions grew from \$880 billion per day to \$3,981 billion. In 2010, 84.9 percent of every currency trade involved the U.S. dollar and 39.1 percent involved the EU's euro (see Table 10.2). Not surprisingly, the U.S. dollar/euro exchange was the most common, with 28 percent of all transactions, followed by the U.S. dollar/yen (14 percent) and the U.S. dollar/UK pound (9 percent).

Note that the total in Table 10.2 is 200 percent rather than 100 percent because every sale is simultaneously a purchase. The dollar is so often traded because it is used as an international medium of exchange and because of the cross-trading that occurs between pairs of currencies. That is, a Chilean importer may pay his or her Mexican supplier in U.S. dollars, or he or she

TABLE 15.2 Composition of Currency Trades, April 2010

Currency	Percent of Total Trades
U.S. dollar	84.9
EU euro	39.1
Japanese yen	19.0
UK pound	12.9
Australian dollar	7.6
Swiss franc	6.4
Canadian dollar	5.3
Other	25.0

Source: Bank for International Settlements, <http://www.bis.org>.

may use Chilean pesos to buy dollars and use the dollars to buy Mexican pesos. It is unlikely that the Mexican exporter would accept Chilean pesos, so one way or another the importer has to come up with dollars.

Currency trading is concentrated in just a few financial centers. London is by far the largest center of foreign exchange trading, as is illustrated by the BIS survey's finding that more U.S. dollars are traded in London than in New York (see Table 10.3). Given the preponderance of the U.S. dollar in currency trades and the importance of London as a trading center, it follows that most of the trades in London do not involve the British pound.

TABLE 15.3 Currency Trading

Centers Location	Percent of World Currency Trading
United Kingdom	36.7
United States	17.9
Japan	6.2
Singapore	5.3
Switzerland	5.2
Hong Kong	4.7
Australia	3.8
France	3.0
Other	17.2

Source: Bank for International Settlements, <http://www.bis.org>.

TABLE 15.4 Major Determinants of an Appreciation or Depreciation

	R Falls: An Appreciation in the Domestic Currency	R Rises: A Depreciation in the Domestic Currency
Long run: Purchasing Power Parity	Home goods are less expensive than foreign goods	Home goods are more expensive than foreign goods
Medium run: The Business Cycle	Domestic economy grows more slowly than foreign	Domestic economy grows faster than foreign
Short run (1): Interest Parity	Home interest rates rise, or foreign rates fall	Home interest rates fall, or foreign rates rise
Short run (2): Speculation	Expectations of a future appreciation	Expectations of a future depreciation

The mechanisms from inconsistent policy to exchange rate crisis and collapse are fairly well understood, but this begs the question about the cause of a sudden shift in expectations. Many recent episodes of sudden exchange rate shifts have occurred when investors lost confidence in a particular currency. Yet why the sudden change in investor confidence? Government policies are often in place for years before they become unsustainable. Quite frequently, an external shock such as a sudden shift in the price of a key input such as oil, or a sudden change in policy by an important trade partner, are the tipping point.

Table 10.4 summarizes the long-, medium-, and short-run factors that have been discussed. The list is not exhaustive, but the main elements are included.

The Real Exchange Rate

The concept of the exchange rate that has been used so far and that is exemplified by the values shown in Table 10.2 does not really tell us what a foreign currency is worth. Exchange rates tell us how many units of domestic currency we give up for one unit of foreign currency, but unless we know what foreign prices are, we still do not know the purchasing power of our domestic money when it is converted to a foreign currency. As an illustration of this problem, suppose that the U.S. dollar–Malaysian ringgit exchange rate is \$0.25 and that it stays constant over the year. However, suppose also that Malaysian inflation is 4 percent while U.S. inflation is 1 percent. After one year, the four ringgits that cost one dollar will buy 3 percent less in Malaysia than the dollar buys in the United States. The relatively higher inflation in Malaysia erodes the value of a dollar's worth of ringgits more rapidly than the dollar loses value at home. Consequently, when converted to ringgits, the real purchasing power of the dollar has declined even though the exchange rate is still \$0.25 per ringgit.

From the point of view of tourists and business people who use foreign exchange, the key item of interest is the purchasing power they get when they

convert their dollars, not the number of units of a foreign currency. An American importer trying to decide between Malaysian and Chinese textiles does not really care if he or she gets four ringgits per dollar or eight Chinese yuan per dollar. The biggest concern is the volume of textiles that can be purchased in Malaysia with four ringgits and in China with eight yuan.

The **real exchange rate** is the market exchange rate (or **nominal exchange rate**) adjusted for price differences. The two are closely connected. By way of illustration, let's consider the case of a wine merchant who is trying to decide whether to stock his or her shop with American or French wine. Let's say that French wine of a given quality costs €200 and American wine of the same quality costs \$180. What the merchant needs to know is the real exchange rate between French and American wine. Suppose that the nominal rate is \$1.20 per euro so that \$180 is equivalent to €150 in the currency market. In this case, French wine costs one-third more than American wine, and the real exchange rate is 1/3 cases of American wine per case of French wine. The algebra is straightforward:

Real exchange rate

$$\begin{aligned} &= [(\text{Nominal exchange rate}) \times (\text{Foreign price})]/(\text{Domestic price}) \\ &= [(\$1.20 \text{ per euro}) \times (\text{€}200 \text{ per case})]/(\$180 \text{ per case}) \\ &= [(\$240 \text{ per case of French wine})]/(\$180 \text{ per case of American wine}) \\ &= 1\frac{1}{3} \text{ cases of American wine per one case of French wine} \end{aligned}$$

Since the real purchasing power of the dollar is much less in France than in the United States, the choice facing the wine merchant is obvious.

In this example, the main lesson is clear. What matters most to exporters and importers is not the nominal exchange rate, but the real exchange rate—in other words, how much purchasing power they have in the countries under comparison. Let R_r symbolize the real exchange rate, R_n the nominal rate. Since we are interested in the whole economy rather than just one market such as the market for wine, we will use a price index to measure overall prices in the two countries. Price indexes are equivalent to the average price of a basket of goods and services in each economy. Let P stand for the home country price index, and P^* represent foreign prices. Then, following the algebra of the wine merchant's calculation:

Real exchange rate

$$= [(\text{Nominal exchange rate}) \times (\text{Foreign prices})]/(\text{Domestic prices}),$$

or, more compactly,

$$R_r = R_n (P^*/P)$$

Suppose, for example, that the U.S. dollar–EU euro nominal exchange rate is \$1.20 per euro and that both price levels are initially set at 100. In this case, the cost of a basket of goods and services is the same in real terms in both countries and

$$R_r = R_n (P^*/P) = R_n (100/100) = R_n$$

The real rate equals the nominal rate when the purchasing power is the same in both countries. Note that purchasing power parity indicates that this is the long-run equilibrium. Over time, however, if inflation is higher at home than in the foreign country, P rises more than P^* , and R_r falls, meaning the domestic currency appreciates in real terms.

By way of illustration, suppose that the United States has 10 percent inflation while the EU has 0 percent. Then, the real U.S.-EU exchange rate (in terms of dollars per euro) would be as follows:

$$R_r = (\$1.20 \text{ per euro}) \times (100/110) = \$1.0909 \text{ per euro}$$

Tourists, investors, and businesspeople can still trade dollars and euros at the nominal rate of \$1.20 per euro (plus whatever commissions they pay to the seller), but the real purchasing power of the U.S. dollar has risen in the EU compared to what it buys at home. The real exchange rate of \$1.0909 per euro tells us that EU goods are now 9 percent cheaper than the U.S. goods that have risen in price. As a result, unless the nominal rate changes, the dollar goes further in the EU than at home. In real terms, the euro has depreciated and the dollar has appreciated.

Changes in the value of real exchange rates play an important role in international macroeconomic relations. When countries control the value of their nominal exchange rate, for example, they must be certain that their prices do not change in relation to the prices of their trading partners. If inflation runs higher at home, then the real value of their currency appreciates. Over a period of time, if uncorrected, this can lead to a build-up in the current account deficit as imports increase and exports decrease. In a number of cases, the end result has been currency crises and the collapse of nominal exchange rates. (For example, Mexico in December 1994 and Thailand in July 1997.)

Alternatives to Flexible Exchange Rates

Fixed exchange rate systems are also called **pegged exchange rate** systems. In these types of systems, there are several possibilities for setting the value of the country's currency. At one extreme, a few (mostly very small) countries give up their currency altogether and adopt the currency of another country, usually the dollar or the euro. More commonly, the value of a nation's money is set equal to a fixed amount of another country's currency, or less commonly to a basket of several currencies. If the exchange rate is not allowed to vary, then it is called a **hard peg**. Fixed exchange rates that fluctuate within a set band are **soft pegs** and these, in turn, can take several forms depending on the amount of variation allowed. Table 10.5 shows that in 2007, there were 23 countries with hard pegs and 82 with soft pegs in which the currency is fixed, but allowed to vary within set limits. Table 10.4 also shows that 83 countries have floating exchange rates. Of these 83 countries, 48 intervene in currency markets when their currencies rise or fall too much in value, while 35 countries let their currencies float independently without intervention.

Through the first seventy years of the twentieth century, fixed exchange rates were the norm, often within a framework that defined the value of a country's

TABLE 15.5 Types of Exchange Rate Systems, 2007

Currency Regime	Countries
Hard pegs	23
Soft pegs	82
Managed floating	48
Independently floating	35
Total	188
More countries have fixed exchange rates than floating.	
Source: IMF, Review of Exchange Arrangements, Restrictions, and Controls, November, 2007.	

currency in terms of a fixed amount of gold. After World War II, many nations shifted away from gold and pegged the value of their currencies to the U.S. dollar or to the currency of another country with which they had strong historical ties. For example, a number of former French colonies in sub-Saharan Africa fixed their currencies to the franc. Beginning in the 1970s, the use of fixed exchange rate systems began a swift decline, first in the high-income industrial economies, and then in many developing countries during the 1980s and 1990s. By the end of the twentieth century, **flexible exchange rate systems** were the norm in every region of the world.

Although the weight of current economic opinion probably favors floating exchange rates, there is widespread recognition that individual country conditions are unique and that there is no single type of exchange rate system appropriate for every country. While the number of countries using flexible exchange rate systems grew rapidly after the early 1970s, it began to decline slightly after 2001. Currently, less than half of the world's nations have flexible rates.

Fixed Exchange Rate Systems

Gold standards are one type of fixed exchange rate that the world abandoned in the 1930s during the Great Depression. Current research shows that the first countries to end the gold standard were the first ones to escape the depression. After World War II, Western economies adopted a modified gold standard under the **Bretton Woods exchange rate system** (1947–1971), but this too was abandoned in the early 1970s. While mainly of historical interest, gold standards are useful to learn about as they highlight a pure form of fixed exchange rate with a hard peg. Under a pure gold standard, nations keep gold as their international reserve. Gold is used to settle most international obligations and nations must be prepared to trade it for their own currency whenever foreigners attempt to “redeem” the home currency they have earned by selling goods and services. In this sense, the nation's money is backed by gold.

There are essentially three rules that countries must follow in order to maintain a gold exchange standard. First, they must fix the value of their currency unit

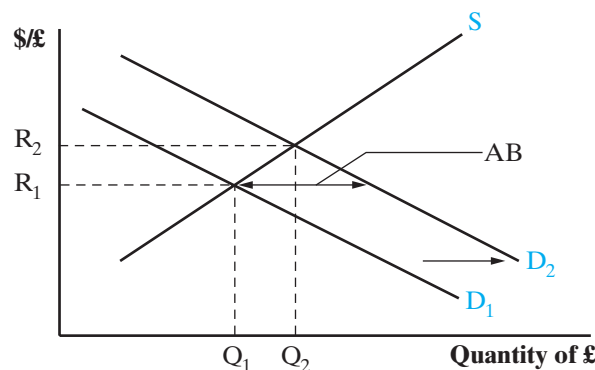
(the dollar, the pound, the yen, and so on) in terms of gold. This fixes the exchange rate. For example, under the modified gold standard of the Bretton Woods exchange rate system, the U.S. dollar was fixed at \$35 per ounce and the British pound was set at £12.5 per ounce. Since both currencies were fixed in terms of gold, they were implicitly set in terms of each other: $\$35 = \text{one ounce of gold} = \text{£}12.5$, or 2.80 dollars per pound ($2.80 = 35/12.5$).

The second rule of the gold standard is that nations keep the supply of their domestic money fixed in some constant proportion to their supply of gold. This requirement is an informal one, but is necessary to ensure that the domestic money supply does not grow beyond the capacity of the gold supply to support it. The third rule of a gold standard is that nations must stand ready and willing to provide gold in exchange for their home country currency.

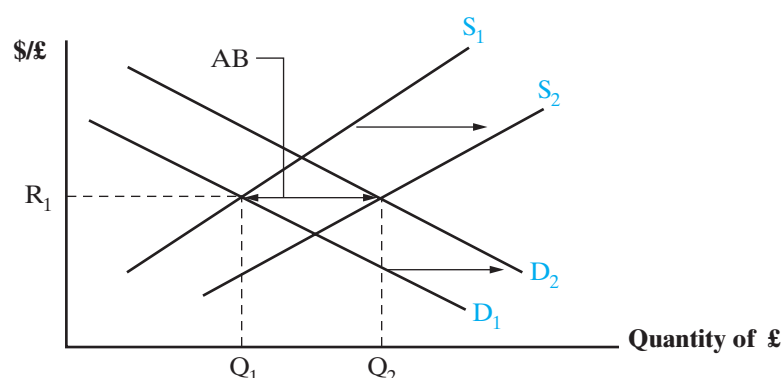
Consider what would happen if a country decided to print large quantities of money for which there is no gold backing. In the short run, purchases of domestically produced goods would rise, causing domestic prices to rise as well. As domestic prices rise, foreign goods become more attractive, since a fixed exchange rate means that they have not increased in price. As imports in the home country increase, foreigners accumulate an unwanted supply of the home country's currency. This is the point at which the gold standard would begin to become unhinged. If gold supplies are low in relation to the supply of domestic currency, the gold reserves will begin to run out at some point as the country pays out gold in exchange for its currency. This spells crisis and a possible end to the gold standard.

Under a fixed exchange rate system, the national supply and demand for foreign currencies may vary but the nominal exchange rate does not. It is the responsibility of the monetary authorities (the central bank or treasury department) to keep the exchange rate fixed. Figure 10.6 illustrates the task before a national

FIGURE 15.6 Fixed Exchange Rates and Changes in Demand



An increase in the demand for British pounds puts pressure on the exchange rate and will cause the dollar to depreciate to R_2 unless the increase in demand is countered by an increase in supply equal to line segment AB.

FIGURE 15.7 Selling Reserves of Pounds to Counter a Weakening Dollar

By selling gold equal in value to AB pounds, the United States prevents a depreciation in the dollar-pound exchange rate.

government when it wishes to keep its currency fixed. Suppose that the United States and the United Kingdom are both on the gold standard and the U.S. demand for British pounds increases.

In the short run or medium run, a rise in demand for pounds from D_1 to D_2 is caused by one of the factors listed in Table 10.4: increased U.S. demand for UK goods, higher UK or lower U.S. interest rates, or speculation that the value of the dollar might not remain fixed for much longer. If R_1 is the fixed U.S.-UK exchange rate, then the United States must counter the weakening dollar and prevent the rate from depreciating to R_2 . One option is to sell the United States' gold reserves in exchange for dollars. This puts gold in the hands of merchants, investors, or speculators who are trying to obtain British pounds. The quantity of gold that must be sold is equivalent to the value of the pounds represented by line segment AB. In effect, the United States meets the increased demand for British pounds by supplying international money—gold—to the market through a sale of some of its gold stock. Since gold and pounds are interchangeable, an increase in the supply of gold is equivalent to an increase in the supply of pounds, as shown in Figure 10.7, and the exchange rate stays at R_1 .

Under a pure gold standard, countries hold gold as a reserve instead of foreign currencies and sell their gold reserves in exchange for their own currency. This action increases the supply of gold—which is international money—and offsets the pressure on the home currency to depreciate. There are two possibilities for the home country as it sells its gold reserves. Either the demand for gold is satisfied and the pressure on its currency eases, or it begins to run out of gold. If the latter happens, the home country may be forced into a devaluation that is accomplished by changing the gold price of its currency. As an illustration, if the dollar is fixed at \$35 per ounce of gold, a devaluation would shift the price of gold to something more than \$35, say \$50, and each ounce of gold sold by the United States buys back a greater quantity of dollars.

Pure gold standards have been rare since the 1930s. More commonly, countries have adopted modified gold standards, such as the Bretton Woods system (see the Case Study), or fixed exchange rate systems called *pegged exchange rates*. Pegged exchange rate systems operate similarly to a gold standard except that instead of gold, another currency is used to “anchor” the value of the home currency.

One potential source of problems with a pegged currency is that the home currency’s value is synchronized with its peg, so changes between the peg and a third-party currency are identical for the home currency and the third party. An example will clarify. Suppose that Thailand decides to peg its currency to the U.S. dollar at the rate of 25 Thai baht per U.S. dollar. The goal of Thailand’s central bank must be to supply dollars whenever it is asked to redeem its own baht. If the dollar appreciates against the Japanese yen, then so does the Thai baht, and at the same rate. Appreciation against the Japanese yen may or may not be a problem for Thailand’s producers, depending on the importance of the Japan-Thailand trade relationship. In 1997, it turned out to be very important, and declining Thai competitiveness from its appreciating currency played a prominent role in triggering the Asian financial crisis of 1997–1998.

The simplest way to avoid this type of problem is to peg the currency not to one single currency, but to a group of currencies. This is, in fact, closer to Thailand’s actual policy in 1997. While this is slightly more complex arithmetically, it reduces the importance of any single country’s currency in the determination of the home country’s currency value. Typically, countries that adopt this strategy select the currencies of their most important trading partners as elements of the basket.

Pegged exchange rates can work very well under many circumstances, but another factor that can cause them to unravel is a significant difference in inflation rates between the home country and its peg. We saw previously that real exchange rates play a greater role in determining trade patterns than nominal rates. Using the United States–Thailand example, and looking at the equation that describes the relationship between real and nominal exchange rates from Thailand’s point of view (as the home country), we have the following:

$$\begin{aligned} R_r &= (25 \text{ baht per dollar}) \times [(U.S. \text{ price level}) / (\text{Thai price level})] \\ &= R_n (P^* / P) \end{aligned}$$

Relatively high inflation in Thailand appears as a faster rate of change in P , and leads to a real appreciation in the baht. Under these circumstances, Thai producers are less competitive and U.S. producers are more so (in Thailand). If the situation persists, speculators will likely step in and begin to sell baht in the expectation that the pegged nominal rate of 25 baht per dollar will be devalued to offset the appreciation in the real rate. Moving the nominal peg from 25 to 30 or 40 baht per dollar may be necessary to restore balance.

The most common technique for dealing with this problem is through the adoption of a **crawling peg**. Crawling pegs are soft pegs that are fixed but periodically adjusted. The idea is to offset any differences in inflation (changes in P) through regular adjustments in R_n . If correctly handled, the real exchange rate

remains constant and the impact of inflation differences never shows up as a change in competitiveness.

There are several other variations on the theme of fixed exchange rates. One of the key points to keep in mind is that purely fixed or purely flexible exchange rate arrangements are rare. When a currency is fixed in value, it is still subject to market pressures of supply and demand, which, at times, can force the government to alter the currency's value. Similarly, when countries adopt a flexible exchange rate system, there is frequently some degree of government intervention in currency markets to try to shape its value.

CASE STUDY

The End of the Bretton Woods System

The Bretton Woods system of exchange rates was enacted at the end of World War II. It included most nations outside the former Soviet Union and its allies. The exchange rate system was a major component of the institutions designed to manage international economic conflict and to support international economic cooperation. In addition to the exchange rate system, the other institutions created at the same time included the International Monetary Fund (IMF), the International Bank for Reconstruction and Development (IBRD) or World Bank, and the General Agreement on Tariffs and Trade (GATT). (See Chapter 2.)

Each institution had its own role in the management of world economic affairs. The roles of the exchange rate were to provide stability by eliminating excess currency fluctuations, to prevent nations from using exchange rate devaluations as a tactic for gaining markets for their goods, and to ensure an adequate supply of internationally accepted reserves so that nations could meet their international obligations.

In the Bretton Woods exchange rate system, the dollar was fixed to gold at the rate of \$1 equaling $\frac{1}{35}$ ounce of gold, or \$35 per ounce. Every other currency within the system was fixed to the dollar and, therefore, indirectly to gold. Unlike a pure gold standard, however, countries could use U.S. dollars as their international reserve and did not have to accumulate gold or tie their money supply to their gold reserves.

The Bretton Woods exchange rate system had one fatal flaw—the dollar. The United States was in a privileged position since its currency was treated the same as gold. This meant that the United States could simply increase its money supply (the supply of dollars) and gain increased purchasing power over European, Japanese, and other countries' goods. Other nations preferred the United States to maintain a relatively robust supply of dollars, since this ensured that there was an adequate supply of international reserves for the world economy.

Problems with this arrangement began when the U.S. economy expanded at a different rate than the economies of its trading partners. In the mid-to-late

1960s, the United States deepened its involvement in the Vietnam War while it simultaneously created the “War on Poverty” at home. Both policies generated large fiscal expenditures that stimulated the economy. While U.S. expansion raced ahead of expansion elsewhere, Europeans found themselves accumulating dollars more rapidly than they desired. The dollars were a by-product of U.S. economic expansion and partially reflected the price increases accompanying the expansion.

Under a different type of exchange rate system, it would have been appropriate for the United States to devalue its currency. U.S. prices had risen relative to foreign prices, the real exchange rate had appreciated as a consequence, and trade deficits were beginning to become a permanent feature of the U.S. economy.

One policy would have been to devalue the nominal dollar exchange rate, but this was not an option. Since every currency was tied to the dollar, there was no way for the United States to devalue against a group of other currencies selectively. An alternative was for the United States to devalue against all currencies by changing the gold value of the dollar. By the late 1960s, it was becoming apparent that this would be necessary.

Persistent U.S. deficits had led to an accumulation of dollars outside the United States, which greatly exceeded the United States’ supply of gold. In other words, the United States lacked the gold reserves to back all of the dollars in circulation. Official recognition of this fact led to the **Smithsonian Agreement** of December 1971, in which the major industrialized countries agreed to devalue the gold content of the dollar by around 8 percent, from \$35 per ounce to \$38.02. In addition, Japan, Germany, and other trade surplus countries increased the value of their currencies.

Although the Smithsonian Agreement was hailed by President Nixon as a fundamental reorganization of international monetary affairs, it quickly proved to be too little and of only temporary benefit. The gold value of the dollar was realigned again in early 1973, from \$38.02 to \$42.22. In addition, further devaluation occurred against other European currencies. The end of the system came in March of 1973 when the major currencies began to float against each other. A few currencies, such as the British pound, had begun to float earlier.

In each case, the strategy of allowing the exchange rate to float in response to supply and demand conditions was adopted as a means of coping with speculation. When speculators had perceived that the dollar was overvalued at \$38 per ounce or \$42 per ounce, they sold dollars in anticipation of a future devaluation. Nor was the dollar the only currency speculated against. Other weak currencies such as the pound and the Italian lira had also been correctly perceived as overvalued and had been sold off by speculators. In the end, the central banks of the weak-currency countries found it impossible to support an unrealistically high value of their currency. The costs of buying up the excess supply of their currencies at overvalued prices proved to be too great. The simplest solution was to let the currencies float.

Choosing the Right Exchange Rate System

Given the menu of choices for exchange rate systems, an active area of economic research has focused on the performance characteristics of systems under different economic conditions and institutional arrangements. For many years, economists debated the pros and cons of fixed and flexible rates, but as the variety of exchange rate options has grown, as capital mobility has increased, and as international trade and investment relations have deepened, researchers have become more concerned with understanding how varying degrees of flexibility or fixity might best serve the interests of individual countries. In particular, economists have tried to learn how different exchange rate systems might influence the core elements of a country's macroeconomy such as the rate of economic growth, the rate of inflation, and the frequency of currency crises.

Traditional views held that countries with fixed exchange rate systems were better at controlling inflation, but that they paid a price in the form of slower economic growth. The reasoning behind this view was that in order to maintain a fixed rate, governments have to be very careful about issuing new money. Since most of the episodes of hyperinflation during the second half of the twentieth century resulted from overexpanding the money supply, it seems reasonable that an exchange rate policy that limits the supply of money would also help avoid inflation. In the view of some economists, however, the limits placed on the ability of a country to manipulate its money supply also remove an important tool that governments use to help manage the rate of economic growth. Therefore, the tradeoff was lower growth for lower inflation.

More recent research, particularly with data from the 1990s, has failed to demonstrate a strong relationship between the type of exchange rate system and either inflation or economic growth. Before the 1990s, countries with fixed or pegged exchange rates tended to have lower rates of inflation, but during the 1990s the differences disappeared. Similarly, there is evidence that countries with more flexible rates tend to have higher average rates of economic growth, but this result depends on the classification of the fastest growing Asian economies. Technically, many of these countries have flexible exchange rates, but at the same time they manage them very closely. When they are omitted from the analysis, there is no significant difference in the rate of growth between countries with relatively fixed and relatively flexible rates. And finally, neither fixed nor flexible rates seem to offer superior protection against a currency crisis. As a result, no particular system seems to rank above any other in its ability to provide superior macroeconomic performance.

Insofar as economists have been able to devise a set of rules for selecting an exchange rate system, they are very general and very basic. If the goal is to find the system that helps minimize negative shocks to an economy, then the source of the shock determines whether a more flexible or more fixed system should be adopted. When the shocks originate in the monetary sector—for example, a central bank that goes overboard in printing new money—a fixed rate is better since it imposes discipline on the central bank. On the other

hand, if the shocks to an economy originate in the external environment—for example, a sudden change in the price of imported oil—then relatively more flexibility in the exchange rate enables the country to adapt to the changes more easily. The general argument here is that individual country characteristics matter a great deal. The problem with these rules, however, is that the source of the shocks to an economy are likely to vary from episode to episode and, as a consequence, the basic rules outlined above provide less practical guidance than desired.

Exchange rate pegs are popular, particularly with many developing countries. There are a couple of reasons for this. First, all economists agree that one of the most important elements of an exchange rate system is its credibility. That is, no matter what type of exchange rate is adopted, a successful system must generate confidence and the widespread belief that it is sustainable. Exchange rate systems that lack credibility are guaranteed to fail in their basic job of providing a smooth and reliable conversion between domestic and foreign money. Under some conditions, exchange rate pegs may offer greater credibility. One of the conditions, and the second reason why some countries continue to peg their currencies, is a relatively high degree of trade dependence on a single, major economy. Consider the case of Mexico, with about 80 percent of its trade with the United States. Because of its trade dependence on the United States, Mexico pegged its peso to the U.S. dollar for many years. Because Mexican inflation ran higher than the U.S. rate, a crawling peg was favored as the means of keeping the real exchange rate relatively constant. The purpose of the dollar peg was to provide benefits to Mexican businesses and consumers by eliminating some of the price variation in Mexican imports and exports. The rule seems to be that when a country is closely tied to the economy of a large, industrial country such as the United States, pegging to its currency may provide additional stability and help businesses plan their futures with greater confidence.

This view is shared by many, but at the same time it is widely accepted that, in Mexico's case, a flexible exchange rate has served it better than the pegged rates it used before 1994. The reason for the discrepancy between what might work in theory and what has worked in practice highlights the complexity of choosing an exchange rate system when every country has unique economic factors and its own set of institutions shaping its economic outcomes. Mexico, due to a set of agreements between the business sector, organized labor, and government, was unable to make the periodic adjustments to its nominal exchange rate that are required with a crawling peg. In effect, Mexico's institutional inability to adjust its nominal exchange rate undermined the credibility of the exchange rate system. The lack of credibility led to periodic bouts of speculation against the peso whenever it was perceived to be overvalued and vulnerable. Several of these speculative bouts were followed by a peso collapse and economic recession. The lesson, in the end, seems to be that the first criterion for choosing an exchange rate system is that it must have credibility in currency and financial markets.

CASE STUDY

Monetary Unions

Some countries prefer not to have their own currency. Seventeen of the twenty-seven countries of the EU use a common currency, the euro, and more are expected to join. Panama adopted the dollar as a legal tender alongside its own currency, called the *cordoba*, in the early twentieth century, and in 2000 Ecuador and El Salvador eliminated their currencies altogether and adopted the dollar.

Dollarization is the term given to the adoption of another country's currency. Dollarization differs from a monetary union, such as the euro zone, because a union has a common central bank that issues the currency and carries out monetary policy. By contrast, the central banks of El Salvador and Ecuador have no ability to issue money, and so they have no control over monetary policy since they cannot expand or contract the money supply. There is no barrier in international law to using another country's money, but in doing so, a country becomes powerless to influence its exchange rate or the quantity of money in circulation.

There are currently four monetary unions in the world. These are the EMU, the Eastern Caribbean Currency Union (ECCU), the West African Economic and Monetary Union (WAEMU), and the Central African Economic and Monetary Community, which is known by its French acronym, CEMAC.

The EU case is discussed in detail in Chapter 14. The two African unions, WAEMU and CEMAC, are the oldest of the monetary unions. Both were formed out of former French colonies in Western Africa and both use the CFA franc as their currency. (CFA stands for *Communauté Française Africaine*, or French African Community.) Both the WAEMU and the CEMAC have central banks that issue their currencies and both fix it to the euro at approximately 655 CFA francs per euro. The French Treasury Department backs both currencies and stands ready to provide currency reserves if either of the two central banks of the monetary unions runs short.

According to most observers, the advantages of CFA francs over independent currencies is that they have lowered inflation in the participating countries and reduced macroeconomic instability. Since the central banks are responsible for more than one economy, it has probably reduced the political influence of individual governments and led to a steadier, less volatile monetary policy. The disadvantages are the same as those for a fixed exchange rate: Changes in the value of the currency cannot be used to protect the domestic economy against shocks that begin outside the country. For example, as the euro gained value against the dollar after 2000 (see Figure 10.1), the CFA franc also appreciated against the dollar and goods produced in the CFA franc zone became more expensive when priced in dollars.

TABLE 15.6 Monetary Unions

Monetary Union	Members	Exchange Rate System
European Monetary Union (EMU)	Seventeen of twenty-seven European Union countries	Flexible
West African Economic and Monetary Union (WAEMU)	Eight countries in sub-Saharan west Africa	Fixed to euro
Central African Economic and Monetary Community (CEMAC)	Six countries in west-central Africa	Fixed to euro
Eastern Caribbean Currency Union (ECCU)	Six island countries and two British territories	Fixed to dollar

This particularly affected the WAEMU, which mainly exports cotton and other agricultural products.

All of the monetary unions are also economic unions (EU), common markets (ECCU is the basis for the Caribbean Common Market), or customs unions (WAEMU and CEMAC). Monetary union implies a high level of integration and coordination and is only worthwhile if other elements of the economy are also integrated. There is not a great deal of agreement as to the value or necessity of monetary unions, but without additional economic integration they make little sense.

Single Currency Areas

On January 1, 1999 eleven members of the European Union adopted the euro as their official currency. As the EU added new members in the first decade of the new century, several chose to use the euro, and by 2011, seventeen of the EU's twenty-seven members had replaced their national currencies with the euro. This was the result of a shared vision of deeper economic, monetary, and political integration that had been developing over many decades. Given that a nation's currency is one of its strongest symbols of national sovereignty, the fact that so many countries have decided to give up their currencies and their ability to set monetary policy is a remarkable set of events.

There are at least four potential reasons why a group of countries might want to share a common currency. First, a single currency eliminates the need to convert each other's money and thereby reduces transaction costs in a number of ways. It eliminates fees paid to the banks or to the currency brokers that arrange the conversion, it simplifies accounting and bookkeeping, and it enables consumers and investors to compare prices across international boundaries more

accurately. Each of these advantages provides some gain in efficiency and a reduction in business costs. Second, a single currency eliminates price fluctuations caused by changes in the exchange rate. When speculators move their money into or out of a country, or when temporary interest rate changes in one country alter the supply and demand for foreign exchange, one country may become (temporarily) cheaper or more expensive for business. As a result, business decisions may reflect temporary shifts in currency values rather than underlying issues of economic efficiency. The elimination of misleading price signals that result from exchange rate fluctuations is also a potential gain in efficiency.

Third, the elimination of exchange rates through the adoption of a single currency can help increase political trust between countries seeking to increase their integration. A single currency removes some of the friction between integrating nations by eliminating the problems that are caused by exchange rate misalignments. Fourth, in some developing countries the adoption of a common currency may give their exchange rate system greater credibility. Use of such a currency can reduce exchange rate fluctuations and create greater confidence in the financial system of the adopting country, possibly leading to lower interest rates and increased availability of credit, although this depends on the overall soundness of the financial system.

Nations that give up their national money do not do so without cost. In addition to its political symbolism, the adoption of a common currency also means that the country no longer has its own money supply as a tool for managing its economic growth. The topic of monetary policy is taken up in more detail in Chapter 11, but the basic point is easy to grasp. Countries with their own currency can influence the rate of growth of the economy in the short run (but not in the long run) through a change in the supply of money. When a country adopts a common currency with one or more other countries, it gives up this tool. After the introduction of the common currency, there is only one money supply and, consequently, one rate of growth of the money supply. New York, for example, shares a common currency with California, and, as a consequence, both states experience the same changes in the money supply. If New York is growing fast and California is growing slowly, it would be impossible for the Federal Reserve to alter the money supply in a way that would speed up growth in California and slow it down in New York. With a single currency, there is a “one-size-fits-all” monetary policy.

Conditions for Adopting a Single Currency

The starting point for analyzing the costs and benefits of a single currency area is the work of Robert Mundell on the theory of **optimal currency areas**. Mundell developed the first set of criteria to determine whether two or more countries are better off sharing a currency instead of using their separate national moneys. Mundell and subsequent research points to four conditions for deciding whether the gains are greater than the costs.

The first condition is that the business cycle must be synchronized and national economies must enter recessions and expansions at more or less the same time. In this case, a single monetary policy is appropriate since each country is individually at the same point in the business cycle and there is no cost associated with the loss of national monetary policy and its replacement with a single policy for all member states. In fact, however, few countries are that well synchronized in their business cycles. Even the states of the United States enter and leave recessions at different points in time, and the national figures on growth only reflect an average across all fifty states.

The second condition is a high degree of labor and capital mobility between the member countries. This allows workers and capital to leave countries or regions where work is scarce and to join the supply of labor and capital in booming regions. In effect, free migration of the factors of production smooths out some of the differences in the business cycle by taking unemployed inputs and moving them to where they are needed. This is how the fifty states of the United States compensate for a lack of complete synchronization in the business cycles of individual states. When conditions are bad in one region, workers and investors move their labor and capital to another region, freeing inputs from areas where they are not needed and providing them to areas where they are.

While capital tends to be relatively mobile, labor is less so, even within countries. Therefore, a third condition is that there are regional policies capable of addressing the imbalances that may develop. Depressed areas may remain depressed if people cannot move or choose not to move because the psychological or other costs are too high or resources outside the region are not available. In the United States, federal taxes and transfer payments help depressed regions adjust and limit some of the shock. When a state is in recession, for example, people still receive their social security checks, Medicare, and other federal transfers. Federal taxes and payments spread the adjustment across the nation and ensure that it is not left up to the state alone. Insofar as the economics of regional policies are concerned, they may be determined at any level, from the currency area (multicountry), to individual nation-states, to subnational units (provinces or cities). The key point is not the agency responsible, but that there are effective policies for assisting regions that may not be synchronized with the majority of the currency area's economy.

Finally, the first three conditions point to the fourth: The nations involved must be seeking a level of integration that goes beyond simple free trade. Free trade requires that nations remove their tariffs, quotas, and other border barriers that inhibit the flow of goods. If this is the goal, a common currency is unnecessary. If something much deeper is sought, however, such as a greater harmonization of national economies and much closer economic and political ties, then a single currency can be helpful, provided the other three conditions are observed. This condition is admittedly ambiguous and is part of the reason why policymakers do not always agree in their analysis. It is somewhat circular reasoning, but true nevertheless, that the desirability of a single currency partly depends on the goals of the countries involved.

CASE STUDY**Is the NAFTA Region an Optimal Currency Area?**

The EU is one model for the creation of a single currency. In the EU model, an entirely new currency is created, and each country gives up its national money. The discussion of a single currency in the NAFTA countries has favored a different model. So far, discussion has centered on the adoption of the U.S. dollar by all three countries instead of the creation of an entirely new currency. Either model leads to the same outcome: a single currency area. Is the discussion realistic? That is, are the proponents of a single currency dreaming or is there something to be gained in such a move?

It is clear that whatever the long-run advantages or disadvantages of a single currency might be, the NAFTA countries have a long way to go before they meet the four conditions necessary for a single currency area to be an optimal policy. First, the business cycles of the three countries have not been synchronized, at least historically. While the macroeconomies of Canada and the United States have often moved together, Mexico has historically had a very different pattern of business cycles. This may be changing, however, since the Mexican cycle appears to be much closer to that of the United States since 1994. Second, given the legal restrictions on labor movement and the political obstacles to opening a North American labor market, labor flows cannot be counted on to help synchronize national business cycles. Third, there are no regional policies within the NAFTA framework and no way to create transfers from one country to another to compensate for slower growth or recession in one area while other places are expanding. Finally, NAFTA was originally conceived as a means to reduce border barriers. While its ultimate goal will surely evolve over time, currently there does not appear to be a consensus that it should be something more than a free-trade area.

As it is now constituted, the NAFTA region is clearly not an optimal currency area. Nevertheless, it is a safe bet that dollarization will continue to be explored, particularly in Mexico. In part, this is because there are counter-arguments to each of the above objections: A single currency will help synchronize the three economies; it is possible to formulate an agreement that allows a guest worker program such as the United States and Mexico had in the 1940s, 1950s, and 1960s; regional policies are simply a matter of political will and financial means, but they would not require huge expenditures; and closer integration of the NAFTA partners is inevitable. Nevertheless, given the problems of the euro zone countries beginning in 2011, it is highly unlikely that any serious analyst might try to push a single currency agenda in the NAFTA region.

Summary



- People hold foreign currency to buy goods and services, to take advantage of interest rate differentials, and to speculate. The primary institutions in the exchange-rate market are commercial banks and foreign exchange brokers.
- Exchange rates can be analyzed with supply and demand analysis, as if they are just another commodity in the economy. Increases (decreases) in the supply of foreign exchange cause the domestic currency to appreciate (depreciate). Increases (decreases) in the demand for foreign exchange cause the domestic currency to depreciate (appreciate).
- Exchange rates are unpredictable because they are simultaneously influenced by long-run, medium-run, and short-run factors. In the long run, purchasing power parity is important. In the medium run, the business cycle is important, and in the short run, interest-rate differentials and speculation are important.
- The interest parity condition says that the interest rate differential between two countries is approximately equal to the percentage difference between the forward and spot exchange rates.
- Firms use forward exchange rate markets to protect against exchange rate risk.
- Real exchange rates are equal to nominal or market exchange rates adjusted for inflation. They give a better picture of the purchasing power of a nation's currency.
- Fixed exchange rate systems were thought to help limit the growth of inflation, but there is little evidence of this over the last two decades. Fixed exchange rates eliminate the ability of governments to use monetary policies to regulate the macroeconomy.
- Flexible exchange rate systems were thought to help increase growth, but there is little evidence of this over the last two decades. Flexible exchange rates free a nation's macroeconomic policies from the need to maintain a fixed exchange rate.
- All exchange rate systems are on a continuum between fixed and flexible rates. Pegged exchange rates, crawling pegs, and a managed float are examples of intermediary-type systems. The most important rule for countries is that their exchange rate system is credible.
- Optimal currency areas are geographical regions within which it is optimal for countries to adopt the same currency. The criteria for an optimal currency area are a synchronized business cycle, complete factor mobility, regional programs for lagging areas, and a desire to achieve a higher level of economic and political integration.

Vocabulary

appreciation	gold standard
Bretton Woods exchange rate system	hard peg
covered interest arbitrage	hedging
crawling peg	interest parity
depreciation	interest rate arbitrage
dollarization	nominal exchange rate
exchange rate	optimal currency area
exchange rate risk	pegged exchange rate
fixed exchange rate system	purchasing power parity
flexible (floating) exchange rate system	real exchange rate
forward exchange rate	Smithsonian Agreement
forward market	soft peg
	spot market

Study Questions

All problems are assignable in [MyEconLab](#): exercises that update with real-time data are marked with .

1. Draw a graph of the supply of and demand for the Canadian dollar by the U.S. market. Diagram the effect of each of the following on the exchange rate; state in words whether the effect is long, medium, or short run; and explain your reasoning.
 - a. More rapid growth in Canada than in the United States
 - b. A rise in U.S. interest rates
 - c. Goods are more expensive in Canada than in the United States
 - d. A recession in the United States
 - e. Expectations of a future depreciation in the Canadian dollar
2. Suppose the U.S. dollar–euro exchange rate is 1.20 dollars per euro, and the U.S.  dollar–Mexican peso rate is 0.10 dollars per peso. What is the euro–peso rate?
3. Suppose the U.S. dollar–yen exchange rate is 0.01 dollars per yen. Since the  base year, inflation has been 2 percent in Japan and 10 percent in the United States. What is the real exchange rate? In real terms, has the dollar appreciated or depreciated against the yen?

4. Which of the three motives for holding foreign exchange are applicable to each of the following?
 - a. A tourist
 - b. A bond trader
 - c. A portfolio manager
 - d. A manufacturer
5. If U.S. visitors to Mexico can buy more goods in Mexico than they can in the United States when they convert their dollars to pesos, is the dollar undervalued or overvalued? Explain.
6. In a fixed exchange rate system, how do countries address the problem of currency market pressures that threaten to lower or raise the value of their currency?
7. In the debate on fixed versus floating exchange rates, the strongest argument for a floating rate is that it frees macroeconomic policy from taking care of the exchange rate. This is also the weakest argument. Explain.
8. Brazil, Argentina, Paraguay, and Uruguay are members of MERCOSUR, a regional trade area that is trying to become a common market. What issues should they consider before they accept or reject a common currency?
9. Suppose that U.S. interest rates are 4 percent more than rates in the EU.
 - a. Would you expect the dollar to appreciate or depreciate against the euro, and by how much?
 - b. If, contrary to your expectations, the forward and spot rates are the same, in which direction would you expect financial capital to flow? Why?
10. Why do some economists claim that the most important feature of any exchange rate system is its credibility?

APPENDIX

The Interest Rate Parity Condition

The following variables are defined as in the chapter:

i = home country interest rate

i^* = foreign interest rate

R = the nominal exchange rate in units of home country currency per unit of foreign currency

F = the forward exchange rate

The forward rate and the interest rates have the same term to maturity.

An investor has a choice between i and i^* . Letting the dollar be the home currency, \$1 invested today will return $\$1(1 + i)$ next period if invested at home. To make the comparison with a foreign investment, the dollar first has to be converted into the foreign currency, then invested, and the earnings must be converted back into dollars. The equivalent of \$1 in foreign currency is $1/R$. If $1/R$ is invested abroad, at the end of the next period it returns $(1/R)(1 + i^*)$, which is in units of foreign currency. The reconversion to dollars can be done in the forward market where the exchange rate for a forward contract is F . Therefore, in dollars, \$1 invested abroad will return $(1/R)(1 + i^*)F$ in the next period.

The interest parity condition states that investors will be indifferent between home and foreign investments (of similar risk), implying that they will move their funds around and cause interest rates and exchange rates to change until the returns are the same in the two cases:

$$1 + i = (1/R)(1 + i^*)F = (1 + i^*)(F/R)$$

Divide by $(1 + i^*)$:

$$(1 + i)/(1 + i^*) = F/R$$

Subtract 1 from both sides:

$$\begin{aligned} [(1 + i)/(1 + i^*)] - [(1 + i^*)/(1 + i^*)] &= F/R - R/R \\ [(1 + i) - (1 + i^*)]/[(1 + i^*)] &= (F - R)/R \\ (i - i^*)/(1 + i^*) &= (F - R)/R \end{aligned}$$

The left-hand side denominator is close to 1 for small values of i^* (this is why we state the interest parity condition as an approximation). The right-hand side is the percentage difference between the forward and spot rates. If it is negative, markets expect an appreciation in the home currency. Rewriting the last equation,

$$i - i^* \approx (F - R)/R,$$

which says that the difference between home country and foreign interest rates is approximately equal to the expected depreciation in the home country currency.



Answers to Selected Exercises

CHAPTER 1

1. a. *False.* There is little or no substitutability across geographical regions of the United States. A consumer in Los Angeles, for example, will not travel to Houston, Atlanta, or New York for lunch just because hamburger prices are lower in those cities. Likewise, a McDonald's or Burger King in New York cannot supply hamburgers in Los Angeles, even if prices were higher in Los Angeles. In other words, a fast-food price increase in New York will affect neither the quantity demanded nor the quantity supplied in Los Angeles or other parts of the country.
- b. *False.* Although consumers are unlikely to travel across the country to buy clothing, suppliers can easily move clothing from one part of the country to another. Thus if clothing prices were substantially higher in Atlanta than Los Angeles, clothing companies could shift supplies to Atlanta, which would reduce the price there.
- c. *False.* Although some consumers might be die-hard Coke or Pepsi loyalists, there are many consumers who will substitute one for the other based on price differences. Thus there is a single market for colas.

CHAPTER 2

2. a. With each price increase of \$20, the quantity demanded decreases by 2. Therefore, $(\Delta Q_D / \Delta P) = -2/20 = -0.1$. At $P = 80$, quantity demanded equals 20 and $E_D = (8/20)(-0.1) = -0.40$. Similarly, at $P = 100$, quantity demanded equals 18 and $E_D = (100/18)(-0.1) = -0.56$.
- b. With each price increase of \$20, quantity supplied increases by 2. Therefore, $(\Delta Q_S / \Delta P) = 2/20 = 0.1$. At $P = 80$, quantity supplied equals 16 and $E_S = (80/16)(0.1) = 0.5$. Similarly, at $P = 100$, quantity supplied equals 18 and $E_S = (100/18)(0.1) = 0.56$.
- c. The equilibrium price and quantity are found where the quantity supplied equals the quantity demanded at the same price. From the table, the $P^* = \$100$ and the $Q^* = 18$ million.
- d. With a price ceiling of \$80, consumers want 20 million, but producers supply only 16 million, for a shortage of 4 million.

3. If Brazil and Indonesia add 200 million bushels of wheat to U.S. wheat demand, the new demand curve will be $Q + 200$, or $Q_D = (3244 - 283P) + 200 = 3444 - 283P$.

Equate supply and the new demand to find the new equilibrium price, $1944 + 207P = 3444 - 283P$, or $490P = 1500$, and thus $P = \$3.06$ per bushel. To find the equilibrium quantity, substitute the price into either the supply or demand equation. Using demand, $Q_D = 3444 - 283(3.06) = 2578$ million bushels.

5. a. Total demand is $Q = 3244 - 283P$; domestic demand is $Q_D = 1700 - 107P$; subtracting domestic demand from total demand gives export demand $Q_E = 1544 - 176P$. The initial market equilibrium price (as given in example) is $P^* = \$2.65$. With a 40-percent decrease in export demand, total demand becomes $Q = Q_D + 0.6Q_E = 1700 - 107P + 0.6(1544 - 176P) = 2626.4 - 212.6P$. Demand is equal to supply. Therefore:

$$2626.4 - 212.6P = 1944 + 207P$$

$$682.4 = 419.6P$$

$$\text{So } P = \frac{682.4}{419.6} = \$1.626 \text{ or } \$1.63. \text{ At this price,}$$

$Q = 2281$. Yes, farmers should be worried. With this drop in quantity and price, revenue goes from \$6609 million to \$3718 million.

- b. If the U.S. government supports a price of \$3.50, the market is not in equilibrium. At this support price, demand is equal to $2626.4 - 212.6(3.5) = 1882.3$ and supply is $1944 + 207(3.5) = 2668.5$. There is excess supply ($2668.5 - 1882.3 = 786.2$) which the government must buy, costing $\$3.50(786.2) = \2751.7 million.
8. a. To derive the new demand curve, we follow the same procedure as in Section 2.6. We know that $E_D = -b(P^*/Q^*)$; substituting $E_D = -0.75$, $P^* = \$3$, and $Q^* = 18$ gives $-0.75 = -b(3/18)$ so that $b = 4.5$. Substituting this value into the equation for the linear demand curve, $Q_D = a - bP$, we have $18 = a - 4.5(3)$. So $a = 31.5$. The new demand curve is $Q_D = 31.5 - 4.5P$.



- b. To determine the effect of a 20-percent decline in copper demand, we note that the quantity demanded is 80 percent of what it would be otherwise for every price. Multiplying the right-hand side of the demand curve by 0.8, $Q_D = (0.8)(31.5 - 4.5P) = 25.2 - 3.6P$. Supply is still $Q_S = -9 + 9P$ and demand is equal to supply. Solving, $P^* = \$2.71$ per pound. A decline in demand of 20 percent, therefore, entails a drop in price of 29 cents per pound, or 9.7 percent.
10. a. First, considering non-OPEC supply: $S_C = Q^* = 19$. With $E_S = 0.05$ and $P^* = \$80$, $E_S = d(P^*/Q)$ implies $d = 0.012$. Substituting for d , $S_C = 19$, and $P = 80$ in the supply equation gives $19 = c + (0.012)(80)$, so that $c = 18.05$. Hence, the supply curve is $S_C = 18.05 + 0.012P$. Similarly, since $Q_D = 32$, $E_D = -b(P^*/Q^*) = -0.05$ and $b = 0.020$. Substituting for b , $Q_D = 32$, and $P = 80$ in the demand equation gives $32 = a - (0.020)(80)$, so that $a = 33.6$. Hence $Q_D = 33.6 - 0.020P$.
- b. The long-run elasticities are: $E_S = 0.30$ and $E_D = -0.30$. As above, $E_S = d(P^*/Q^*)$ and $E_D = -b(P^*/Q^*)$, implying $0.30 = d(80/19)$ and $-0.30 = -b(80/32)$. So $d = 0.07$ and $b = -0.12$. Next solve for c and a : $S_C = c + dP$ and $Q_D = a - bP$, which implies that $19 = c + (0.07)(80)$ and $32 = a - (0.12)(80)$. Therefore, $c = 13.3$ and $a = 41.6$.
- c. The discovery of new oil fields will increase OPEC supply by 2bb/yr, so $S_C = 19$, $S_O = 15$, and $D = 34$. The new short-run total supply curve is $S_T = 33.05 + 0.012P$. Demand is unchanged: $D = 33.6 - 0.020P$. Since supply equals demand, $33.05 + 0.012P = 33.6 - 0.020P$. Solving, $P = \$17.19$ per barrel. An increase in OPEC supply entails a drop in price of \$62.81, or 79% in the short-run.
- To analyze the long-run, use the new long-run supply curve, $S_T = 28.3 + 0.071P$. Setting this equal to long-run demand gives: $28.3 + 0.071P = 41.6 - 0.120P$, so that $P = \$69.63$ per barrel, only \$10.37 per barrel (13%) less than the original long-run price.
- CHAPTER 3
1. a. In free-market equilibrium, $L^S = L^D$. Solving, $w = \$4$ and $L^S = L^D = 40$. If the minimum wage is \$5, then $L^S = 50$ and $L^D = 30$. The number of people employed will be given by the labor demand. So employers will hire 30 million workers.
- b. With the subsidy, only $w - 1$ is paid by the firm. The labor demand becomes $L^{D*} = 80 - 10(w - 1)$. So $w = \$4.50$ and $L = 45$.
4. a. Equating demand and supply, $28 - 2P = 4 + 4P - P^* = 4$ and $Q^* = 20$.
- b. The 25-percent reduction would imply that farmers produce 15 billion bushels. To encourage farmers to withdraw their land from cultivation, the government must give them 5 billion bushels that they can sell on the market. Since the total supply to the market is still 20 billion bushels, the market price remains at \$4 per bushel. Farmers gain because they incur no costs for the 5 billion bushels received from the government. We calculate these cost savings by taking the area under the supply curve between 15 and 20 billion bushels. The prices when $Q = 15$ and when $Q = 20$ are $P = \$2.75$ and $P = \$4.00$. The total cost of producing the last 5 billion bushels is therefore the area of a trapezoid with a base of $20 - 15 = 5$ billion and an average height of $(2.75 + 4.00)/2 = 3.375$. The area is $5(3.375) = \$16.875$ billion.
- c. Taxpayers gain because the government does not have to pay to store the wheat for a year and then ship it to an underdeveloped country. The PIK Program can last only as long as wheat reserves last. But PIK assumes that the land removed from production can be restored to production at such time as the stockpiles are exhausted. If this cannot be done, consumers may eventually pay more for wheat-based products. Finally, farmers enjoy a windfall profit because they have no production costs.
10. a. To find the price of natural gas when the price of oil is \$60 per barrel, equate the quantity demanded and quantity supplied of natural gas, and solve for P_G . The relevant equations are: Supply: $Q = 15.90 + 0.72P_G + 0.05P_O$, Demand: $Q = 0.02 - 1.8P_G + 0.69P_O$. Using $P_O = \$60$, we get: $15.90 + 0.72P_G + 0.05(60) = 0.02 - 1.8P_G + 0.69(60)$, so the price of natural gas is $P_G = \$8.94$. Substituting into the supply or the demand curve gives a free-market quantity of 25.34 Tcf. If a maximum price of natural gas were set at \$3, the quantity supplied would be 21.06 Tcf and the quantity demanded would be 36.02 Tcf. To calculate the deadweight loss, we measure the area of triangles B and C (see Figure 3.4). To find area B we must first determine the price on the demand curve when quantity equals 21.1. From the demand equation, $21.1 = 41.42 - 1.8P_G$. Therefore, $P_G = \$11.29$. Area B equals $(0.5)(25.3 - 21.1)(11.29 - 8.94) = \4.9 billion, and area C is $(0.5)(25.3 - 21.1)(8.94 - 3) = \12.5 billion. The deadweight loss is $4.9 + 12.5 = \$17.4$ billion.
- b. To find the price of oil that would yield a free market price of natural gas of \$3, we set the quantity demanded equal to the quantity supplied, use $P_G = \$3$, and solve for P_O . Therefore, $Q_S = 15.90 + 0.72(3) + 0.05P_O = 0.02 - 1.8(3) + 0.69P_O = Q_D$, or $18.06 + 0.05P_O = -5.38 + 0.69P_O$, so that $0.64P_O = 23.44$ and $P_O = \$36.63$. This yields a free market price of natural gas of \$3.

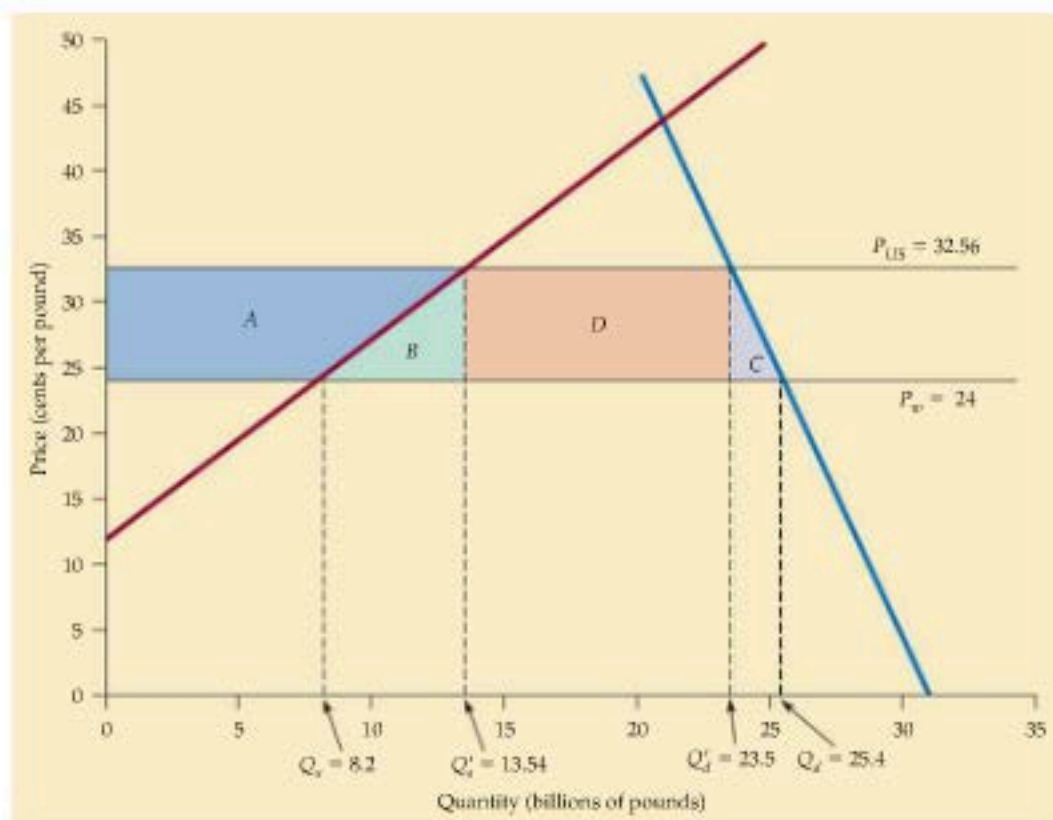


FIGURE 3(a)

11. a. To find the new domestic price, we set the quantity demanded minus the quantity supplied equal to 10. Therefore, $Q_D - Q_S = (29.73 - 0.19P) - (-7.95 + 0.66P) = 10$. $0.85P = 27.68$, meaning that $P =$

32.56 cents. If imports had been expanded to 10 billion pounds, the U.S. price would have fallen by 3.44 cents.

- b. Substituting the new price of 32.56 cents into the supply and demand equations, we find that the U.S. production of sugar would decrease to 13.54 billion pounds, while demand would increase to 23.54 billion pounds, with the additional 10 billion pounds supplied by imports. In order to find the change in the consumer and producer surpluses, it might help to redraw the graph as Figure 3(a). The gain to producers is given by the area of trapezoid A: $A = \left(\frac{1}{2} \times (32.56 - 24)(8.2)\right) + (13.54 - 8.2)(32.56 - 24) = \930 million, which is \$500 million less than the producer gain when imports were limited to 6.9 billion pounds.

To find the gain to consumers, we must find the change in the lost consumer surplus, given by the sum of trapezoid A, triangles B and C, and rectangle D. We've already found the area of trapezoid A. Triangle $B = \frac{1}{2}(32.56 - 24)(13.54 - 8.2) = \228.52 million, triangle $C = \frac{1}{2}(32.56 - 24)(25.4 - 23.54) = \79.47 million, and rectangle $D = (32.56 - 24)(23.54 - 13.54) = \856.34 million. The sum of A, B, C, and D is \$2.09 billion. When imports were limited to 6.9 billion pounds, the loss in consumer surplus

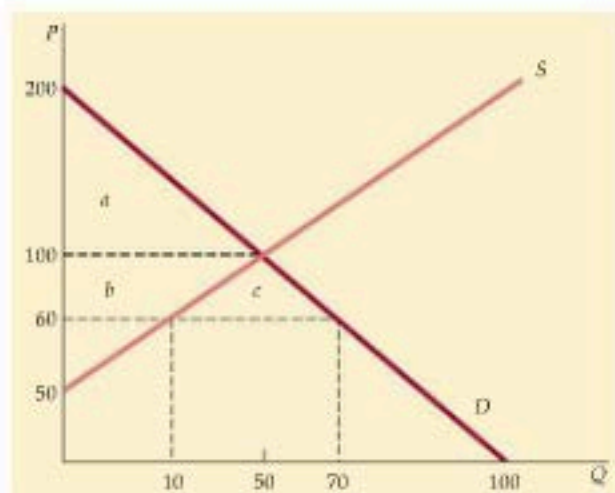


FIGURE 3(b)



is \$2.88 billion, meaning that consumers gain about \$790 million when imports are raised to 10 billion pounds.

- c. The deadweight loss is given by the sum of the areas of triangles B and C: $B = \frac{1}{2}(32.56 - 24)(13.54 - 8.2) = \228.52 million and $C = \frac{1}{2}(32.56 - 24)(25.4 - 23.54) = \79.47 million. $B + C = \$228.52 + \$79.47 = \$308$ million. To find the change in deadweight loss from Example 9.6, we subtract this from the original deadweight loss of \$614.22 million. $\$614.22 \text{ million} - \$308 \text{ million} = \$306.22 \text{ million}$. In other words, raising the import quota to 10 billion pounds per year reduces the deadweight loss by \$306.22 million.

The gain to foreign producers is given by the area of rectangle D. When imports are limited to 6.9 billion pounds, $D = \$836.4$ million; when imports are raised to 10 billion pounds, $D = (32.56 - 24)(23.54 - 13.54) = \856.34 million. Because the U.S. price of sugar has increased, foreign producers are able to earn higher profits – about \$19.94 million, to be exact.

12. First, equate supply and demand to determine equilibrium quantity: $50 + Q = 200 - 2Q$, or $Q_{EQ} = 50$ (million pounds). Substitute $Q_{EQ} = 50$ into either the supply or demand equation to determine price: $P_s = 50 + 50 = 100$ and $P_d = 200 - (2)(50) = 100$. Thus, the equilibrium price P is \$1 (100 cents). However, the world market price is 60 cents. At this price, the domestic quantity supplied is $60 = 50 + Q_s$, or $Q_s = 10$, and domestic demand is $60 = 200 - 2Q_d$, or $Q_d = 70$. Imports equal the difference between domestic demand and supply,

or 60 million pounds. If Congress imposes a tariff of 40 cents, the effective price of imports increases to \$1. At \$1, domestic producers satisfy domestic demand and imports fall to zero.

As shown in Figure 3(b), consumer surplus before the tariff is equal to area $a + b + c$, or $(0.5)(200 + 60)(70) = 4,900$ million cents or \$49 million. After the tariff, the price rises to \$1.00 and consumer surplus falls to area a , or $(0.5)(200 - 100)(50) = \$25$ million, a loss of \$24 million. Producer surplus increases by area b , or $(100 - 60)(10) + (0.5)(100 - 60)(50 - 10) = \12 million. Finally, because domestic production is equal to domestic demand at \$1, no hula beans are imported and the government receives no revenue. The difference between the loss of consumer surplus and the increase in producer surplus is deadweight loss which is \$12 million.

13. No, they would not. The clearest case is where labor markets are competitive. With either design of the tax, the wedge between supply and demand must total 12.4 percent of the wage paid. It does not matter whether the tax is imposed entirely on the workers (shifting the effective supply curve up by 12.4 percent) or entirely on the employers (shifting the effective demand curve down by 12.4 percent). The same applies to any combination of the two that sums to 12.4 percent.

CHAPTER 7

2. The four mutually exclusive states are given in Table 7 below.

TABLE 7		
	CONGRESS PASSES TARIFF	CONGRESS DOES NOT PASS TARIFF
Slow growth rate	State 1: Slow growth with tariff	State 2: Slow growth without tariff
Fast growth rate	State 3: Fast growth with tariff	State 4: Fast growth without tariff

4. The expected value is $EV = (0.4)(100) + (0.3)(30) + (0.3)(-30) = \40 . The variance is $\sigma^2 = (0.4)(100 - 40)^2 + (0.3)(30 - 40)^2 + (0.3)(-30 - 40)^2 = 2,940$.
8. Initially, total wealth is \$450,000. We calculate expected wealth under three options. Under the safe option, $E(U) = (450,000 + 1.05^2 200,000)^{1/2} = 678$.

With the summer corn crop, $E(U) = .7(250,000 + 500,000)^{1/2} + .3(250,000 + 50,000)^{1/2} = 770$. Finally, with the drought resistant summer corn crop, $E(U) = .7(250,000 + 450,000)^{1/2} + .3(250,000 + 350,000)^{1/2} = 818$. The option with the highest expected utility is planting the drought resistant crop.

12. To determine the total demand curve, we add up 100 standard demand curves and 100 rule of thumb demand curves: $Q = 100*(20 - P) + 100*(10 \text{ if } P < 10 \text{ or } 0 \text{ if } P \geq 10) = 3000 - 100P$ if $P < 10$ and $2000 - 100P$ if $P \geq 10$. The resulting total demand curve is given to the right.

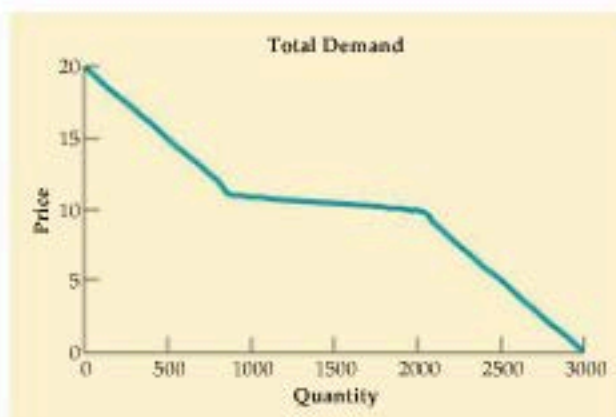


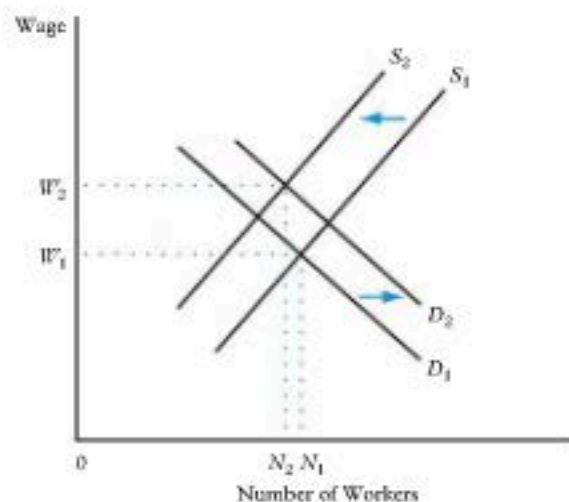
FIGURE 7

Answers to Odd-Numbered Review Questions and Problems

Chapter 4

Review Questions

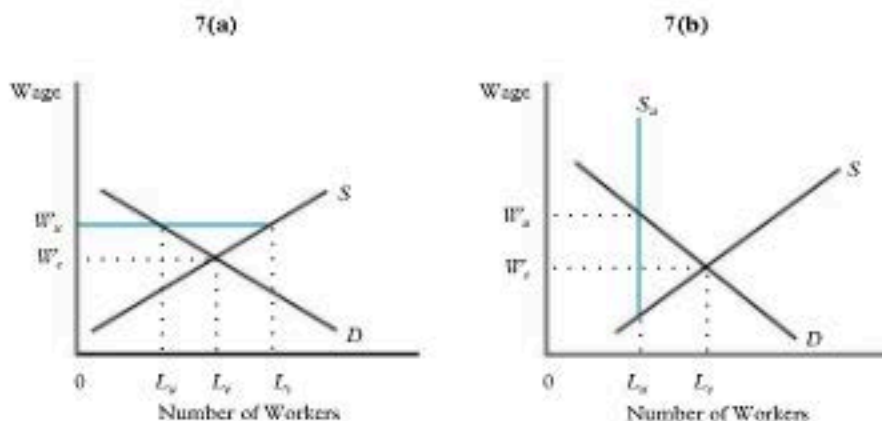
1. As shown in the figure, the outflow of construction workers shifted the labor supply curve relevant to Egypt's construction sector to the *left* (from S_1 to S_2), while the demand curve for the services of construction workers shifted to the *right* (D_1 to D_2). Because both shifts, by themselves, tended to increase the equilibrium wage rate from W_1 to W_2 , we would clearly expect wages in the Egyptian construction sector to have risen faster than average. However, the two shifts by themselves had opposite effects on employment, so the expected net change in employment is theoretically ambiguous.



3. Many engineers are employed in research and development tasks. Therefore, if a major demander of research and development were to reduce its demand, the demand curve for engineers would shift left, causing their wages and employment to fall.
5. If the wages for arc welders are above the equilibrium wage, the company is paying more for its arc welders than it needs to and, as a result, is hiring fewer than it could. Thus, the definition of overpayment that makes the most sense in this case is one in which the wage rate is above the equilibrium wage.

A ready indicator of an above-equilibrium wage rate is a long queue of applicants whenever a position in a company becomes available. Another indicator is an abnormally low quit rate as workers (in this case arc welders) who are lucky enough to obtain the above-equilibrium wage cling tenaciously to their jobs.

7.



9. This regulation essentially increases the cost of capital and will have an ambiguous effect on the demand curve for labor. On the one hand, the increased cost of capital will increase the cost of production and cause a scale effect that tends to depress employment. On the other hand, this regulation will increase the cost of capital relative to labor and could stimulate the substitution of labor for capital. Thus, the substitution effect will work to increase employment while the scale effect will work to decrease it. Which effect is stronger cannot be predicted from theory alone.
11. a. Economic growth tends to shift the labor demand curve to the right (more workers are demanded at each wage rate).
- b. Greater job growth accompanied by slower positive wage changes will result if the labor supply curve in Canada is flatter (has a smaller positive slope) than the labor supply curve in the United States.

Problems

1. Unemployment rate = $100 \times (\text{number unemployed}) / (\text{number unemployed} + \text{number employed}) = 100 \times (5 \text{ million}) / (135 \text{ million}) = 3.7 \text{ percent}$. Labor force participation rate = $100 \times (\text{number employed} + \text{number unemployed}) / \text{adult population} = 100 \times (135 \text{ million}) / (210 \text{ million}) = 64.3 \text{ percent}$.
3. The quickest places to find the relevant data are probably at <http://www.bls.gov/ces/>, "Tables from Employment and Earnings" (Table B-11), and <http://www.bls.gov>, Consumer Price Index. If average hourly earnings are rising faster than the Consumer Price Index (CPI), then real wages have been rising. In addition, we should consider the impact of mismeasurement in the CPI. If the CPI overstates inflation (as discussed in the text), then real wages have risen more rapidly than the official statistics suggest. The Bureau of Labor Statistics Web site contains links to recent research on changes in the construction of the CPI that are intended to remove some of the historical bias.

5. Real hourly minimum wage in 1990 = nominal wage in 1990 / CPI in 1990
 $= (\$3.80 / 130.7) \times 100$
 $= \$2.91$

$$\begin{aligned}\text{Real hourly minimum wage in 2006} &= \text{nominal wage in 2006} / \text{CPI in 2006} \\ &= (\$5.15 / 201.6) \times 100 \\ &= \$2.55\end{aligned}$$

The federal minimum wage decreased in real dollars from 1990 to 2006.

7. If cashiers are being paid \$8.00 per hour, they are being paid more than the market equilibrium wage for their job. At \$8.00 per hour, employers will hire 110 cashiers, but 175 workers are available for work as a cashier. There are 65 workers who would like a job as a cashier at a wage of \$8.00 per hour but cannot get such a job. Because a labor surplus exists for jobs that are overpaid, a wage above equilibrium has two implications. First, employers are paying more than necessary to produce their output; they could cut wages and still find enough qualified workers for their job openings. In fact, if they did cut wages, they could expand output and make their product cheaper and more accessible to consumers. Second, more workers want jobs than can find them. If wages were reduced a bit, more of these disappointed workers could find work.

Chapter 5

Review Questions

1. Profit maximization requires that firms hire labor until marginal revenue productivity equals the market wage. If wages are low, a profit maximizer will hire labor in abundant quantity, driving the marginal revenue productivity down to the low level of the wage. This statement, then, seems to imply that firms are not maximizing profits.
3. The potential employment effects of OSHA standards differ with the type of approach taken. If the standards apply to capital (machinery), they will increase the cost of capital equipment. This increase in cost has a scale effect, which will reduce the quantity demanded of all inputs (including labor). On the other hand, it also provides employers with an incentive to substitute labor (which is now relatively cheaper) for capital in producing any given desired level of output. This substitution will moderate the decline in employment.

In contrast, requiring employers to furnish personal protective devices to employees increases the cost of labor. In this case, employers have an incentive to substitute now relatively cheaper capital for labor when producing any given level of output (as above, the increased cost of production causes a scale effect that also tends to reduce employment).

Other things equal, then, the employment reduction induced by safety standards will be greater if the personal protective device method is used. However, to fully answer the question requires information on the costs of meeting the standards using the two methods. For example, if the "capital" approach increases capital costs by 50 percent while the "personal protective" approach increases labor costs by only 1 percent, the scale effect in the first method will probably be large enough that greater employment loss will be associated with the first method.

5. The wage and employment effects in both service industries and manufacturing industries must be considered. In the service sector, the wage tax on employers can be analyzed in much the same way that payroll taxes are analyzed in the text. That is, a tax on wages, collected from the employer, will cause the demand curve to shift leftward if the curve is drawn with respect to the wage that employees take home. At any given hourly wage that employees take home, the cost to the employer has risen by the amount of the tax. An increase in cost associated with any employee wage dampens the employer's appetite for labor and causes the demand curve to shift down and to the left.

The effects on employment and wages depend on the shape of the labor supply curve. If the labor supply curve is upward-sloping, both employment and the wage employees take home will fall. If the supply curve is vertical, employment will not fall, but wages will fall by the full amount of the tax. If the supply curve is horizontal, the wage rate will not fall, but employment will.

The reduced employment and/or wages in the service sector should cause the supply of labor to the manufacturing sector to shift to the right (as people formerly employed in the service sector seek employment elsewhere). This shift in the supply curve should cause employment in manufacturing to increase even if the demand curve there remains stationary. If the demand curve does remain stationary, the employment increase would be accompanied by a decrease in manufacturing wages. However, the demand for labor in manufacturing may also shift to the right as consumers substitute away from the now more expensive services and buy the now relatively cheaper manufactured goods. If this demand shift occurs, the increase in employment would be accompanied by either a wage increase or a smaller wage reduction than would occur if the demand curve for labor in manufacturing were to remain stationary.

7. The imposition of financial penalties on employers who are discovered to have hired illegal immigrants essentially raises the cost of hiring them. The employers now must pay whatever the prevailing wage of the immigrants is, and they also face the possibility of a fine if they are discovered to have illegally employed workers. This penalty can be viewed as increasing the cost of hiring illegal workers so that this cost now exceeds the wage. This effect can be seen as a leftward shift of the demand curve for illegal immigrants, thus reducing their employment and wages.

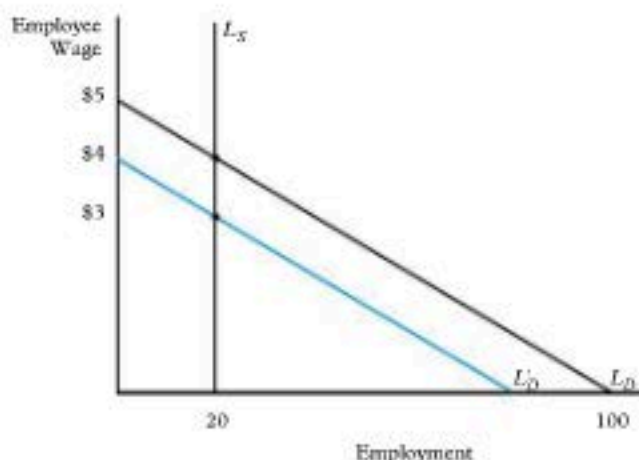
The effects on the demand for skilled “natives” depend on whether skilled and unskilled labor are gross substitutes or gross complements. Raising the cost of unskilled labor produces a scale effect that tends to increase the cost of production and reduce skilled employment. If skilled and unskilled labor are complements in production, the demand for skilled labor will clearly shift to the left as a result of the government’s policy. However, if they are substitutes in production, the increased costs of unskilled labor would stimulate the substitution of skilled for unskilled labor. In this case, the demand for skilled labor could shift either right (if the substitution effect dominated the scale effect) or left (if the scale effect dominated).

9. Wage subsidies shift the demand for labor curve (in terms of employee wages) to the right. The effect on employment depends on the slope of the labor supply curve, which affects how much of the increased demand is

translated into wage increases. The increases in employment will be greater when the supply curve is flatter and the associated wage increase received by workers is smaller.

Problems

1. The marginal product (as measured by these test scores) is 0.
3. See the figure below. Since the supply curve is vertical, the workers will bear the entire tax. The wage will fall by \$1 per hour, from \$4 to \$3.



5. (Appendix) As the chapter explains, to minimize cost, the firm picks K and L so that $W/MP_L = C/MP_K$, where C is the rental cost of capital. Rearrange this $W/C = MP_L/MP_K$ and substitute in the information from the problem:

$$12/4 = 30K^{0.25}L^{-0.25} / 10K^{-0.75}L^{0.75}$$

$$3 = 3K/L$$

$$K = L$$

7. a. Pick K and L so that $(MP_L/MP_K = W/C$ or $(25K/25L) = 8/8 = 1$.
b. Since the cost-minimizing capital-labor ratio is 1, the firm should use equal amounts of capital and labor. To produce 10,000 pairs of earrings, the calculation is as follows:

$$Q = 25K \times L$$

$$10,000 = 25K \times L$$

$$400 = K \times L$$

Since $K = L$, $400 = K \times K$. 20 units of both K and L must be used, and at \$8 per unit the cost comes to \$320.

- c. Costs are minimized when $MP_L/MP_K = W/C$. MP_L equals $25K$, and MP_K equals $25L$, so their ratio equals K/L . For costs to be minimized, K/L must now equal $8/6$, meaning that the capital-labor ratio rises from 1 to 1.33. Once capital becomes cheaper, capital is substituted for labor.

Chapter 6

Review Questions

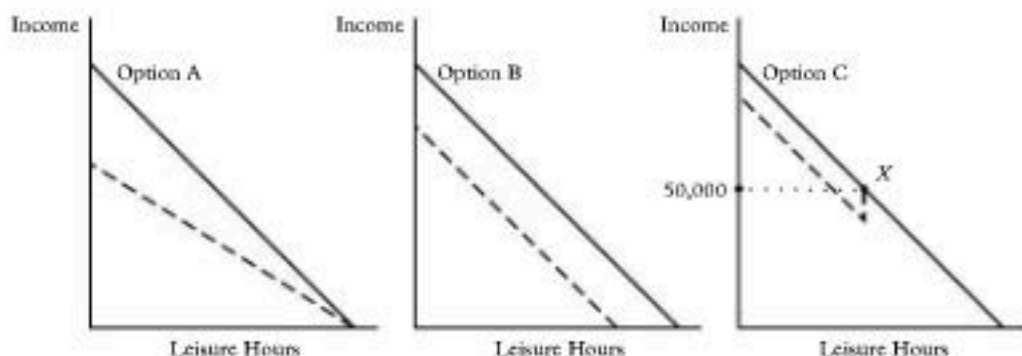
- False. An inferior good is defined as one that people consume less of as their incomes rise (if the price of the good remains constant). A labor supply curve is drawn with respect to a person's wage rate. Thus, for a labor supply curve to be backward-bending, the supply curve must be positively sloped in some range and then become negatively sloped in another. A typical way of illustrating a backward-bending supply curve is shown below.



Along the positively sloped section of this backward-bending supply curve, the substitution effect of a wage increase dominates the income effect, and as wages rise, the person increases his or her labor supply. However, after the wage reaches W_0 in the figure, further increases in the wage are accompanied by a reduction in labor supply. In this negatively sloped portion of the supply curve, the income effect dominates the substitution effect.

We have assumed that the income effect is negative and that, therefore, leisure is a normal good. Had we assumed leisure to be an inferior good, the increases in wealth brought about by increased wages would have worked *with* the underlying substitution effect and caused the labor supply curve to be unambiguously positively sloped.

3. The graphs for each option are shown below, with the new constraints shown as dashed lines. By mandating that 5 percent of each hour be worked for free, option A reduces lawyers' wages, creating income and substitution effects that work in opposite directions on their desired labor supply.



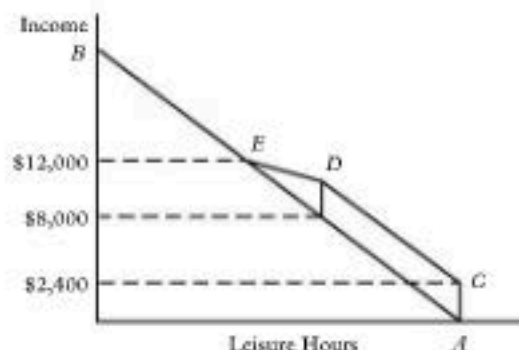
Option B essentially reduces the time lawyers have available for leisure and paid work, which shifts the budget constraint to the left in a parallel manner (keeping the wage rate constant). This creates an income effect that increases their incentives to work for pay.

Option C leaves unchanged the budget constraint of lawyers who work relatively few hours, but for those who work enough to earn over \$50,000, there is an income effect that tends to increase work incentives. For some whose incomes were only slightly above \$50,000, however, the \$5,000 tax may drive them to reduce hours of work, thereby reducing their earnings to \$50,000 and avoiding the tax. These lawyers find their utilities are maximized at point X in the graph of option C's budget constraint.

5. Absenteeism is one dimension of labor supply, so the proposals must be analyzed using labor supply theory. Both proposals increase worker income, because employees now have paid sick days; this increase in income will tend to increase absenteeism through the income effect. The first proposal also raises the *hourly wage*, however, because any unused sick leave can be converted to cash in direct proportion to the unused days. Thus, this first proposal will tend to have a substitution effect accompanying the income effect, so the overall expected change in absenteeism is ambiguous.

The second proposal raises the cost of the *first* sick day because, if absent, the worker loses the entire promised insurance policy. Thus, there is a huge substitution effect offsetting the income effect for the first day of absence. However, once sick leave is used at all, *further* days of absence cause no further loss of pay; thus, after the first day, there is no substitution effect to offset the income effect, and this will tend to increase the incentives for absenteeism.

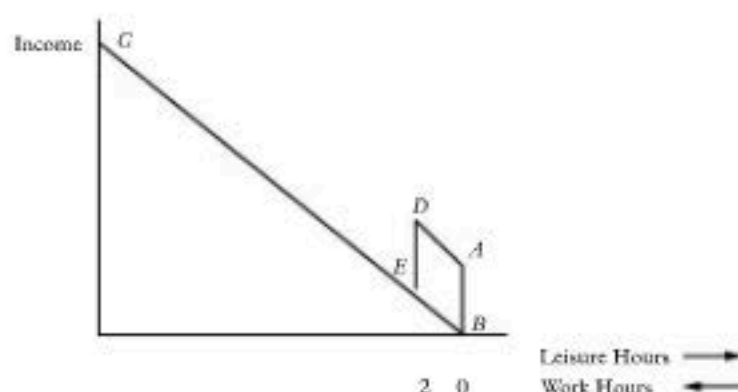
7. In the following figure, the straight line AB represents the person's market constraint (that is, the constraint in a world with no subsidies). $ACDEB$ is the constraint that would apply if the housing subsidy proposal became effective.



The effects on labor supply depend on which segment of $ACDEB$ the person finds relevant. There are four possible cases. First, if the indifference curves are very steeply pitched (reflecting a strong desire to consume leisure), the housing subsidy proposal will not affect work incentives. The person strongly desiring leisure would continue to not work (would be at point C) but would receive the housing subsidy of \$2,400. The second case occurs when the person has a tangency along segment CD . Along this segment, the person's effective wage rate is the same as the market wage, so there is a pure income effect tending to reduce work incentives.

If the person has a tangency point along segment DE , there are likewise reduced incentives to work because the income effect caused by the northeast shifting out of the budget constraint is accompanied by a reduction in the effective wage rate. Finally, those with tangency points along EB will not qualify for the housing subsidy program and therefore will not alter their labor supply behavior. (An exception to this case occurs when a person with a tangency point near point E before the initiation of the housing subsidy program now has a tangency point along segment DE and, of course, works less than before.)

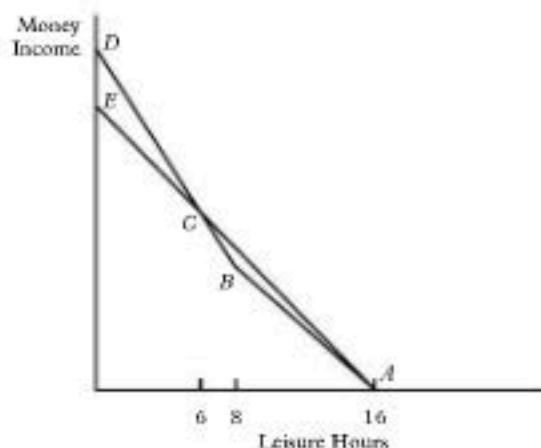
9. The old constraint is ABC below, and the new one is $BADEC$.



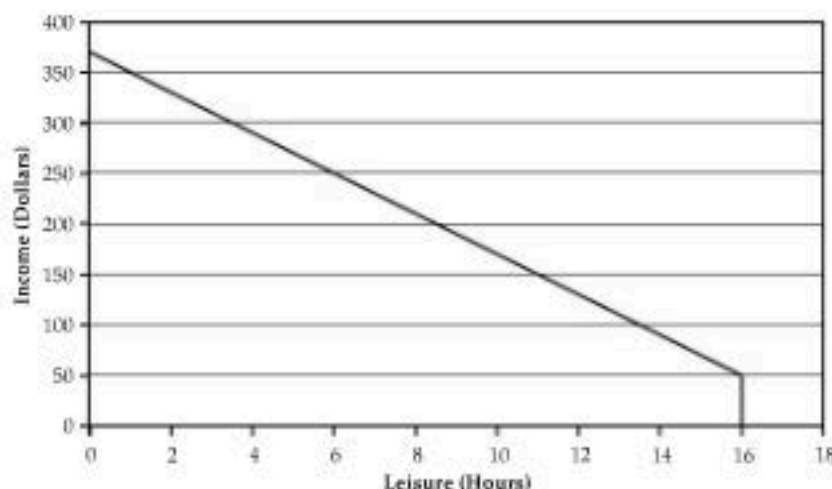
The work-incentive effects of the new constraint depend on worker preferences. For those with relatively strong preferences for leisure, who may not have been in the labor force before, there is an increased incentive to join the labor force and work part-time. For workers with very weak preferences for leisure (who had a tangency along the upper part of EC before), there will be no effect. However, for workers whose earlier tangency was along the lower middle part of EC , the new constraint may create incentives to cut the hours of work and maximize utility at point D .

Problems

1. a. See the following figure, where the initial budget constraint is given by ACE . After the new law is passed, the budget constraint bends upward after 8 hours of work. Thus, the new wage rate and overtime constraint is given by $ABCD$, which intersects the old constraint at point C —the original combination of income and working hours (10 hours of work in this example).



- b. Initially, earnings were $\$11 \times 10 = \110 . The new earnings formula is $8W + 2 \times 1.5W$, where W = the hourly wage. Pick W so that this total equals $\$110$. Since $11W = \$110$, we calculate that $W = \$10$ per hour.
- c. See the figure above. If the workers were initially at a point of utility maximization, their initial indifference curve was tangent to the initial budget constraint (line ACE) at point C . Since the new budget constraint (along segment BD) has a steeper slope ($\$15$ per hour rather than $\$11$ per hour), the workers' initial indifference curve cannot be tangent to the new constraint at point C . Instead, there will be a new point of tangency along segment CD , and hours of work must increase—tangency points along CD lie to the left of point C . (Income in the vicinity of point C is effectively being held constant, and the substitution effect always pulls in the direction of less leisure whenever the wage rate has risen.)
3. a. Δ hours worked per year = Δ hours worked per week \times weeks worked per year = $(-10)(50) = -500$
 Income Effect = $(\Delta H / \Delta Y) | W(\text{constant}) = -500 / 50000 = -1/100$
 Interpretation: For every $\$100$ increase in nonlabor income, you work 1 hour less each year.
- b. The substitution effect is zero. The lottery win enhances wealth (income) independent of the hours of work. Thus, income is increased without a change in the compensation received from an hour of work.
- 5.



7. Teddy's nonlabor income is $\$75$. His base wage rate is $(\$145 - \$75) / (16 - 9) = 70/7 = \10 per hour. His overtime wage rate is $(\$325 - \$145) / (9 - 0) = \$180/9 = \20 per hour. Teddy needs to work at least 7 hours before he receives overtime.

Photo Credits



CHAPTER 1

p. 9 Natsuki Sakai/AFLO/Newscom

CHAPTER 2

p. 26 Joegough/Dreamstime

p. 27 Vladislav Gajic/Shutterstock

p. 35 Orientaly/Shutterstock

p. 44 Bajinda/Shutterstock

p. 52 Slavoljub Pantelic/Shutterstock

CHAPTER 3

p. 71 Antonia Reeve/Photo Researchers, Inc.

p. 76 Michael Rosa/Shutterstock

p. 81 David Frazier/Corbis

p. 88 Tim Page/Corbis

p. 95 Heather A. Craig/Shutterstock

CHAPTER 7

p. 227 Amy Etra/PhotoEdit, Inc.

p. 229 Antonia Reeve/Photo Researchers, Inc.

p. 237 Gerald Holubowicz/Alamy

p. 240 Robyn Beck/AFP/Getty Images/
Newscom

p. 242 Corbis/SuperStock

p. 250 Quavondo/iStockphoto

Chapters 1, 2, 3, and 7 are taken from *Microeconomics*, Eighth Edition by Robert S. Pindyck and Daniel L. Rubinfeld.

Chapters 4, 5, and 6 are taken from *Modern Labor Economics: Theory and Public Policy*, Twelfth Edition by Ronald G. Ehrenberg and Robert S. Smith

Chapters 8 through 12 are taken from *The Economics of Money, Banking, and Financial Markets*, Fifth Canadian Edition by Frederic S. Mishkin and Apostolos Serletis.

Chapters 13, 14, and 15 are taken from *International Economics*, Sixth Edition by James Gerber.

Chapter 1: Preliminaries

1. Decide whether each of the following statements is true or false and explain why:

- a. **Fast food chains like McDonald's, Burger King, and Wendy's operate all over the United States. Therefore the market for fast food is a national market.**

This statement is false. People generally buy fast food locally and do not travel large distances across the United States just to buy a cheaper fast food meal. Because there is little potential for arbitrage between fast food restaurants that are located some distance from each other, there are likely to be multiple fast food markets across the country.

- b. **People generally buy clothing in the city in which they live. Therefore there is a clothing market in, say, Atlanta that is distinct from the clothing market in Los Angeles.**

This statement is false. Although consumers are unlikely to travel across the country to buy clothing, they can purchase many items online. In this way, clothing retailers in different cities compete with each other and with online stores such as L.L. Bean. Also, suppliers can easily move clothing from one part of the country to another. Thus, if clothing is more expensive in Atlanta than Los Angeles, clothing companies can shift supplies to Atlanta, which would reduce the price in Atlanta. Occasionally, there may be a market for a specific clothing item in a faraway market that results in a great opportunity for arbitrage, such as the market for blue jeans in the old Soviet Union.

- c. **Some consumers strongly prefer Pepsi and some strongly prefer Coke. Therefore there is no single market for colas.**

This statement is false. Although some people have strong preferences for a particular brand of cola, the different brands are similar enough that they constitute one market. There are consumers who do not have strong preferences for one type of cola, and there are consumers who may have a preference, but who will also be influenced by price. Given these possibilities, the price of cola drinks will not tend to differ by very much, particularly for Coke and Pepsi.

3. **At the time this book went to print, the minimum wage was \$7.25. To find the current value of the *CPI*, go to <http://www.bls.gov/CPI/home.htm>. Click on "*CPI* Tables," which is found on the left side of the web page. Then, click on "Table Containing History of *CPI*-U.S. All Items Indexes and Annual Percent Changes from 1913 to Present." This will give you the *CPI* from 1913 to the present.**

- a. **With these values, calculate the current real minimum wage in 1990 dollars.**

The last year of data available when these answers were prepared was 2010. Thus, all calculations are as of 2010. You should update these values for the current year.

$$\text{Real minimum wage in 2010} = \frac{CPI_{1990}}{CPI_{2010}} \times \text{minimum wage in 2010} = \frac{130.7}{218.056} \times \$7.25 = \$4.35.$$

So, as of 2010, the real minimum wage in 1990 dollars was \$4.35.

- b. **Stated in real 1990 dollars, what is the percentage change in the real minimum wage from 1985 to the present?**

The minimum wage in 1985 was \$3.35. You can get a complete listing of historical minimum wage rates from the Department of Labor, Wage and Hour Division at <http://www.dol.gov/whd/minwage/chart.htm>.

$$\text{Real minimum wage in 1985} = \frac{CPI_{1990}}{CPI_{1985}} \times \$3.35 = \frac{130.7}{107.6} \times \$3.35 = \$4.07.$$

The real minimum wage therefore increased slightly from \$4.07 in 1985 to \$4.35 in 2010 (all in 1990 dollars). This was an increase of $\$4.35 - \$4.07 = \$0.28$, so the percentage change was $(0.28/4.07) \times 100\% = 6.88\%$.

Chapter 2: The Basics of Supply and Demand

1. Suppose the demand curve for a product is given by $Q = 300 - 2P + 4I$, where I is average income measured in thousands of dollars. The supply curve is $Q = 3P - 50$.

- a. If $I = 25$, find the market-clearing price and quantity for the product.

Given $I = 25$, the demand curve becomes $Q = 300 - 2P + 4(25)$, or $Q = 400 - 2P$. Set demand equal to supply and solve for P and then Q :

$$400 - 2P = 3P - 50$$

$$P = 90$$

$$Q = 400 - 2(90) = 220.$$

- b. If $I = 50$, find the market-clearing price and quantity for the product.

Given $I = 50$, the demand curve becomes $Q = 300 - 2P + 4(50)$, or $Q = 500 - 2P$. Setting demand equal to supply, solve for P and then Q :

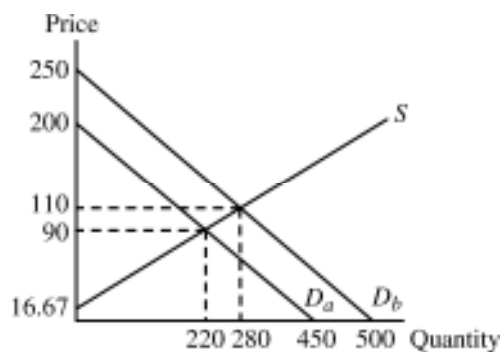
$$500 - 2P = 3P - 50$$

$$P = 110$$

$$Q = 500 - 2(110) = 280.$$

- c. Draw a graph to illustrate your answers.

It is easier to draw the demand and supply curves if you first solve for the inverse demand and supply functions, i.e., solve the functions for P . Demand in part a is $P = 200 - 0.5Q$ and supply is $P = 16.67 + 0.333Q$. These are shown on the graph as D_a and S . Equilibrium price and quantity are found at the intersection of these demand and supply curves. When the income level increases in part b, the demand curve shifts up and to the right. Inverse demand is $P = 250 - 0.5Q$ and is labeled D_b . The intersection of the new demand curve and original supply curve is the new equilibrium point.



3. Refer to Example 2.5 (page 37) on the market for wheat. In 1998, the total demand for U.S. wheat was $Q = 3244 - 283P$ and the domestic supply was $Q_S = 1944 + 207P$. At the end of 1998, both Brazil and Indonesia opened their wheat markets to U.S. farmers. Suppose that these new markets add 200 million bushels to U.S. wheat demand. What will be the free-market price of wheat and what quantity will be produced and sold by U.S. farmers?

If Brazil and Indonesia add 200 million bushels of wheat to U.S. wheat demand, the new demand curve will be $Q + 200$, or

$$Q_D = (3244 - 283P) + 200 = 3444 - 283P.$$

Equate supply and the new demand to find the new equilibrium price.

$$1944 + 207P = 3444 - 283P, \text{ or}$$

$$490P = 1500, \text{ and thus } P = \$3.06 \text{ per bushel.}$$

To find the equilibrium quantity, substitute the price into either the supply or demand equation. Using demand,

$$Q_D = 3444 - 283(3.06) = 2578 \text{ million bushels.}$$

5. **Much of the demand for U.S. agricultural output has come from other countries. In 1998, the total demand for wheat was $Q = 3244 - 283P$. Of this, total domestic demand was $Q_D = 1700 - 107P$, and domestic supply was $Q_S = 1944 + 207P$. Suppose the export demand for wheat falls by 40%.**

- a. **U.S. farmers are concerned about this drop in export demand. What happens to the free-market price of wheat in the United States? Do farmers have much reason to worry?**

Before the drop in export demand, the market equilibrium price is found by setting total demand equal to domestic supply:

$$3244 - 283P = 1944 + 207P, \text{ or}$$

$$P = \$2.65.$$

Export demand is the difference between total demand and domestic demand: $Q = 3244 - 283P$ minus $Q_D = 1700 - 107P$. So export demand is originally $Q_e = 1544 - 176P$. After the 40% drop, export demand is only 60% of the original export demand. The new export demand is therefore, $Q'_e = 0.6Q_e = 0.6(1544 - 176P) = 926.4 - 105.6P$. Graphically, export demand has pivoted inward as illustrated in the figure below.

The new total demand becomes

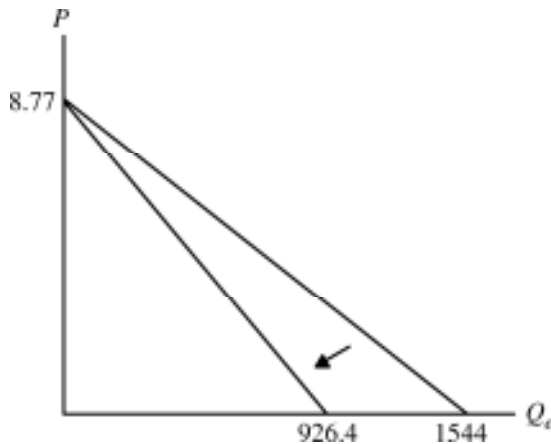
$$Q' = Q_D + Q'_e = (1700 - 107P) + (926.4 - 105.6P) = 2626.4 - 212.6P.$$

Equating total supply and the new total demand,

$$1944 + 207P = 2626.4 - 212.6P, \text{ or}$$

$$P = \$1.63,$$

which is a significant drop from the original market-clearing price of \$2.65 per bushel. At this price, the market-clearing quantity is about $Q = 2281$ million bushels. Total revenue has decreased from about \$6609 million to \$3718 million, so farmers have a lot to worry about.



- b. Now suppose the U.S. government wants to buy enough wheat to raise the price to \$3.50 per bushel. With the drop in export demand, how much wheat would the government have to buy? How much would this cost the government?

With a price of \$3.50, the market is not in equilibrium. Quantity demanded and supplied are

$$Q' = 2626.4 - 212.6(3.50) = 1882.3, \text{ and}$$

$$Q_s = 1944 + 207(3.50) = 2668.5.$$

Excess supply is therefore $2668.5 - 1882.3 = 786.2$ million bushels. The government must purchase this amount to support a price of \$3.50, and will have to spend $\$3.50(786.2 \text{ million}) = \2751.7 million.

7. In 2010, Americans smoked 315 billion cigarettes, or 15.75 billion packs of cigarettes. The average retail price (including taxes) was about \$5.00 per pack. Statistical studies have shown that the price elasticity of demand is -0.4 , and the price elasticity of supply is 0.5 .

- a. Using this information, derive linear demand and supply curves for the cigarette market.

Let the demand curve be of the form $Q = a - bP$ and the supply curve be of the form $Q = c + dP$, where a , b , c , and d are positive constants. To begin, recall the formula for the price elasticity of demand

$$E_P^D = \frac{P}{Q} \frac{\Delta Q}{\Delta P}.$$

We know the demand elasticity is -0.4 , $P = 5$, and $Q = 15.75$, which means we can solve for the slope, $-b$, which is $\Delta Q/\Delta P$ in the above formula.

$$\begin{aligned} -0.4 &= \frac{5}{15.75} \frac{\Delta Q}{\Delta P} \\ \frac{\Delta Q}{\Delta P} &= -0.4 \left(\frac{15.75}{5} \right) = -1.26 = -b. \end{aligned}$$

To find the constant a , substitute for Q , P , and b in the demand function to get $15.75 = a - 1.26(5)$, so $a = 22.05$. The equation for demand is therefore $Q = 22.05 - 1.26P$. To find the supply curve, recall the formula for the elasticity of supply and follow the same method as above:

$$E_p^S = \frac{P}{Q} \frac{\Delta Q}{\Delta P}$$

$$0.5 = \frac{5}{15.75} \frac{\Delta Q}{\Delta P}$$

$$\frac{\Delta Q}{\Delta P} = 0.5 \left(\frac{15.75}{5} \right) = 1.575 = d.$$

To find the constant c , substitute for Q , P , and d in the supply function to get $15.75 = c + 1.575(5)$ and $c = 7.875$. The equation for supply is therefore $Q = 7.875 + 1.575P$.

- b. **In 1998, Americans smoked 23.5 billion packs cigarettes, and the retail price was about \$2.00 per pack. The decline in cigarette consumption from 1998 to 2010 was due in part to greater public awareness of the health hazards from smoking, but was also due in part to the increase in price. Suppose that the *entire decline* was due to the increase in price. What could you deduce from that about the price elasticity of demand?**

Calculate the arc elasticity of demand since we have a range of prices rather than a single price. The arc elasticity formula is

$$E_p = \frac{\Delta Q}{\Delta P} \frac{\bar{P}}{\bar{Q}}$$

where \bar{P} and \bar{Q} are average price and quantity, respectively. The change in quantity was $15.75 - 23.5 = -7.75$, and the change in price was $5 - 2 = 3$. The average price was $(2 + 5)/2 = 3.50$, and the average quantity was $(23.5 + 15.75)/2 = 19.625$. Therefore, the price elasticity of demand, assuming that the *entire decline* in quantity was due solely to the price increase, was

$$E_p = \frac{\Delta Q}{\Delta P} \frac{\bar{P}}{\bar{Q}} = \frac{-7.75}{3} \frac{3.50}{19.625} = -0.46.$$

9. **In Example 2.8 (page 52), we discussed the recent increase in world demand for copper, due in part to China's rising consumption.**

- a. **Using the original elasticities of demand and supply (i.e., $E_S = 1.5$ and $E_D = -0.5$), calculate the effect of a 20% increase in copper demand on the price of copper.**

The original demand is $Q = 27 - 3P$ and supply is $Q = -9 + 9P$ as shown on page 51. The 20% increase in demand means that the new demand is 120% of the original demand, so the new demand is $Q'_D = 1.2Q$. $Q'_D = (1.2)(27 - 3P) = 32.4 - 3.6P$. The new equilibrium is where Q'_D equals the original supply:

$$32.4 - 3.6P = -9 + 9P.$$

The new equilibrium price is $P^* = \$3.29$ per pound. An increase in demand of 20%, therefore, entails an increase in price of 29 cents per pound, or 9.7%.

- b. **Now calculate the effect of this increase in demand on the equilibrium quantity, Q^* .**

Using the new price of \$3.29 in the supply curve, the new equilibrium quantity is $Q^* = -9 + 9(3.29) = 20.61$ million metric tons per year, an increase of 2.61 million metric tons (mmt) per year. Except for rounding, you get the same result by plugging the new price of \$3.29 into the new demand curve. So an increase in demand of 20% entails an increase in quantity of 2.61 mmt per year, or 14.5%.

- c. As we discussed in Example 2.8, the U.S. production of copper declined between 2000 and 2003. Calculate the effect on the equilibrium price and quantity of *both* a 20% increase in copper demand (as you just did in part a) *and* of a 20% decline in copper supply.

The new supply of copper falls (shifts to the left) to 80% of the original, so $Q'_S = 0.8Q = (0.8)(-9 + 9P) = -7.2 + 7.2P$. The new equilibrium is where $Q'_D = Q'_S$.

$$32.4 - 3.6P = -7.2 + 7.2P$$

The new equilibrium price is $P^* = \$3.67$ per pound. Plugging this price into the new supply equation, the new equilibrium quantity is $Q^* = -7.2 + 7.2(3.67) = 19.22$ million metric tons per year. Except for rounding, you get the same result if you substitute the new price into the new demand equation. The combined effect of a 20% increase in demand and a 20% decrease in supply is that price increases by 67 cents per pound, or 22%, and quantity increases by 1.22 mmt per year, or 6.8%, compared to the original equilibrium.

11. Refer to Example 2.10 (page 59), which analyzes the effects of price controls on natural gas.

- a. Using the data in the example, show that the following supply and demand curves describe the market for natural gas in 2005–2007:

$$\text{Supply: } Q = 15.90 + 0.72P_G + 0.05P_O$$

$$\text{Demand: } Q = 0.02 - 1.8P_G + 0.69P_O$$

Also, verify that if the price of oil is \$50, these curves imply a free-market price of \$6.40 for natural gas.

To solve this problem, apply the analysis of Section 2.6 using the definition of cross-price elasticity of demand given in Section 2.4. For example, the cross-price elasticity of demand for natural gas with respect to the price of oil is:

$$E_{GO} = \left(\frac{\Delta Q_G}{\Delta P_O} \right) \left(\frac{P_O}{Q_G} \right).$$

$\left(\frac{\Delta Q_G}{\Delta P_O} \right)$ is the change in the quantity of natural gas demanded because of a small change in the price of oil, and for linear demand equations, it is constant. If we represent demand as

$Q_G = a - bP_G + eP_O$ (notice that income is held constant), then $\left(\frac{\Delta Q_G}{\Delta P_O} \right) = e$. Substituting this into

the cross-price elasticity, $E_{GO} = e \left(\frac{P_O^*}{Q_G^*} \right)$, where P_O^* and Q_G^* are the equilibrium price and quantity.

We know that $P_O^* = \$50$ and $Q_G^* = 23$ trillion cubic feet (Tcf). Solving for e ,

$$1.5 = e \left(\frac{50}{23} \right), \text{ or } e = 0.69.$$

Similarly, representing the supply equation as $Q_G = c + dP_G + gP_O$, the cross-price elasticity of supply is $g \left(\frac{P_O^*}{Q_G^*} \right)$, which we know to be 0.1. Solving for g , $0.1 = g \left(\frac{50}{23} \right)$, or $g = 0.5$ rounded to one decimal place.

We know that $E_S = 0.2$, $P_G^* = 6.40$, and $Q^* = 23$. Therefore, $0.2 = d\left(\frac{6.40}{23}\right)$, or $d = 0.72$. Also,

$E_D = -0.5$, so $-0.5 = -b\left(\frac{6.40}{23}\right)$, and thus $b = 1.8$.

By substituting these values for d , g , b , and e into our linear supply and demand equations, we may solve for c and a :

$23 = c + 0.72(6.40) + 0.05(50)$, so $c = 15.9$, and

$23 = a - 1.8(6.40) + 0.69(50)$, so that $a = 0.02$.

Therefore, the supply and demand curves for natural gas are as given. If the price of oil is \$50, these curves imply a free-market price of \$6.40 for natural gas as shown below. Substitute the price of oil in the supply and demand equations. Then set supply equal to demand and solve for the price of gas.

$$15.9 + 0.72P_G + 0.05(50) = 0.02 - 1.8P_G + 0.69(50)$$

$$18.4 + 0.72P_G = 34.52 - 1.8P_G$$

$$P_G = \$6.40.$$

- b. Suppose the regulated price of gas were \$4.50 per thousand cubic feet instead of \$3.00. How much excess demand would there have been?**

With a regulated price of \$4.50 for natural gas and the price of oil equal to \$50 per barrel,

$$\text{Demand: } Q_D = 0.02 - 1.8(4.50) + 0.69(50) = 26.4, \text{ and}$$

$$\text{Supply: } Q_S = 15.9 + 0.72(4.50) + 0.05(50) = 21.6$$

With a demand of 26.4 Tcf and a supply of 21.6 Tcf, there would be an excess demand (i.e., a shortage) of 4.8 Tcf.

- c. Suppose that the market for natural gas remained unregulated. If the price of oil had increased from \$50 to \$100, what would have happened to the free-market price of natural gas?**

In this case

$$\text{Demand: } Q_D = 0.02 - 1.8P_G + 0.69(100) = 69.02 - 1.8P_G, \text{ and}$$

$$\text{Supply: } Q_S = 15.9 + 0.72P_G + 0.05(100) = 20.9 + 0.72P_G.$$

Equating supply and demand and solving for the equilibrium price,

$$20.9 + 0.72P_G = 69.02 - 1.8P_G, \text{ or } P_G = \$19.10.$$

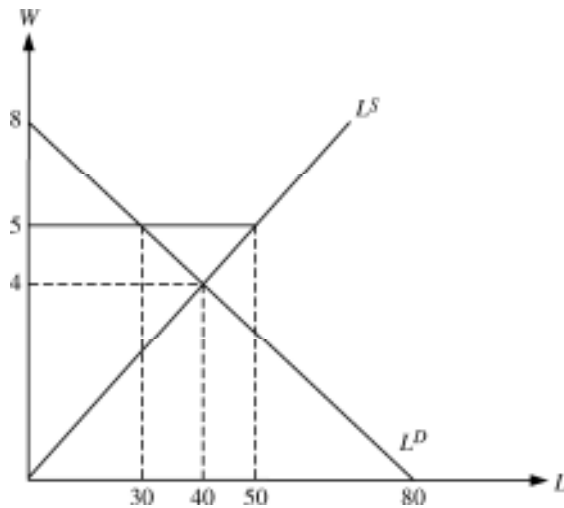
The free-market price of natural gas would have almost tripled from \$6.40 to \$19.10.

Chapter 3: The Analysis of Competitive Markets

1. From time to time, Congress has raised the minimum wage. Some people suggested that a government subsidy could help employers finance the higher wage. This exercise examines the economics of a minimum wage and wage subsidies. Suppose the supply of low-skilled labor is given by $L^S = 10w$, where L^S is the quantity of low-skilled labor (in millions of persons employed each year), and w is the wage rate (in dollars per hour). The demand for labor is given by $L^D = 80 - 10w$.

- a. What will be the free-market wage rate and employment level? Suppose the government sets a minimum wage of \$5 per hour. How many people would then be employed?

In a free-market equilibrium, $L^S = L^D$. Solving yields $w = \$4$ and $L^S = L^D = 40$. If the minimum wage is \$5, then $L^S = 50$ and $L^D = 30$. The number of people employed will be given by the labor demand, so employers will hire only 30 million workers.



- b. Suppose that instead of a minimum wage, the government pays a subsidy of \$1 per hour for each employee. What will the total level of employment be now? What will the equilibrium wage rate be?

Let w_s denote the wage received by the sellers (i.e., the employees), and w_b the wage paid by the buyers (the firms). The new equilibrium occurs where the vertical difference between the supply and demand curves is \$1 (the amount of the subsidy). This point can be found where

$$L^D(w_b) = L^S(w_s), \text{ and}$$

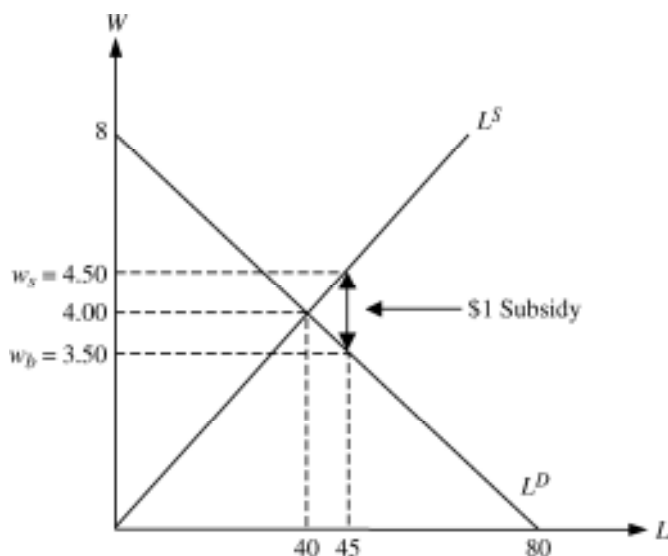
$$w_s - w_b = 1.$$

Write the second equation as $w_b = w_s - 1$. This reflects the fact that firms pay \$1 less than the wage received by workers because of the subsidy. Substitute for w_b in the demand equation:

$$L^D(w_b) = 80 - 10(w_s - 1), \text{ so}$$

$$L^D(w_b) = 90 - 10w_s.$$

Note that this is equivalent to an upward shift in demand by the amount of the \$1 subsidy. Now set the new demand equal to supply: $90 - 10w_s = 10w_s$. Therefore, $w_s = \$4.50$, and $L^D = 90 - 10(4.50) = 45$. Employment increases to 45 (compared to 30 with the minimum wage), but wage drops to \$4.50 (compared to \$5.00 with the minimum wage). The net wage the firm pays falls to \$3.50 due to the subsidy.

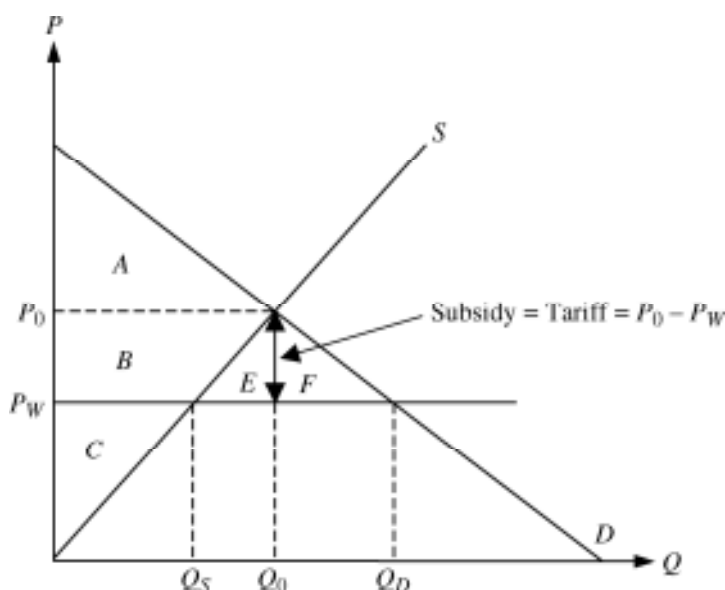


3. Japanese rice producers have extremely high production costs, due in part to the high opportunity cost of land and to their inability to take advantage of economies of large-scale production. Analyze two policies intended to maintain Japanese rice production: (1) a per-pound subsidy to farmers for each pound of rice produced, or (2) a per-pound tariff on imported rice. Illustrate with supply-and-demand diagrams the equilibrium price and quantity, domestic rice production, government revenue or deficit, and deadweight loss from each policy. Which policy is the Japanese government likely to prefer? Which policy are Japanese farmers likely to prefer?

We have to make some assumptions to answer this question. If you make different assumptions, you may get different answers. Assume that initially the Japanese rice market is open, meaning that foreign producers and domestic (Japanese) producers both sell rice to Japanese consumers. The

world price of rice is P_W . This price is below P_0 , which is the equilibrium price that would occur in the Japanese market if no imports were allowed. In the diagram below, S is the domestic supply, D is the domestic demand, and Q_0 is the equilibrium quantity that would prevail if no imports were allowed. The horizontal line at P_W is the world supply of rice, which is assumed to be perfectly elastic. Initially Japanese consumers purchase Q_D rice at the world price. Japanese farmers supply Q_S at that price, and $Q_D - Q_S$ is imported from foreign producers.

Now suppose the Japanese government pays a subsidy to Japanese farmers equal to the difference between P_0 and P_W . Then Japanese farmers would sell rice on the open market for P_W plus receive the subsidy of $P_0 - P_W$. Adding these together, the total amount Japanese farmers would receive is P_0 per pound of rice. At this price they would supply Q_0 pounds of rice. Consumers would still pay P_W and buy Q_D . Foreign suppliers would import $Q_D - Q_0$ pounds of rice. This policy would cost the government $(P_0 - P_W)Q_0$, which is the subsidy per pound times the number of pounds supplied by Japanese farmers. It is represented on the diagram as areas $B + E$. Producer surplus increases from area C to $C + B$, so $\Delta PS = B$. Consumer surplus is not affected and remains as area $A + B + E + F$. Deadweight loss is area E , which is the cost of the subsidy minus the gain in producer surplus.



Instead, suppose the government imposes a tariff rather than paying a subsidy. Let the tariff be the same size as the subsidy, $P_0 - P_W$. Now foreign firms importing rice into Japan will have to sell at the world price plus the tariff: $P_W + (P_0 - P_W) = P_0$. But at this price, Japanese farmers will supply Q_0 , which is exactly the amount Japanese consumers wish to purchase. Therefore there will be no imports, and the government will not collect any revenue from the tariff. The increase in producer surplus equals area B , as it is in the case of the subsidy. Consumer surplus is area A , which is less than it is under the subsidy because consumers pay more (P_0) and consume less (Q_0). Consumer surplus decreases by $B + E + F$. Deadweight loss is $E + F$: the difference between the decrease in consumer surplus and the increase in producer surplus.

Under the assumptions made here, it seems likely that producers would not have a strong preference for either the subsidy or the tariff, because the increase in producer surplus is the same under both policies. The government might prefer the tariff because it does not require any government expenditure. On the other hand, the tariff causes a decrease in consumer surplus, and government officials who are elected by consumers might want to avoid that. Note that if the subsidy and tariff

amounts were smaller than assumed above, some tariffs would be collected, but we would still get the same basic results.

5. About 100 million pounds of jelly beans are consumed in the United States each year, and the price has been about 50 cents per pound. However, jelly bean producers feel that their incomes are too low and have convinced the government that price supports are in order. The government will therefore buy up as many jelly beans as necessary to keep the price at \$1 per pound. However, government economists are worried about the impact of this program because they have no estimates of the elasticities of jelly bean demand or supply.

- a. Could this program cost the government *more* than \$50 million per year? Under what conditions? Could it cost *less* than \$50 million per year? Under what conditions? Illustrate with a diagram.

If the quantities demanded and supplied are very responsive to price changes, then a government program that doubles the price of jelly beans could easily cost more than \$50 million. In this case, the change in price will cause a large change in quantity supplied, and a large change in quantity demanded. In Figure 9.5.a.i, the cost of the program is $(\$1)(Q_S - Q_D)$. If $Q_S - Q_D$ is larger than 50 million, then the government will pay more than \$50 million. If instead supply and demand are relatively inelastic, then the increase in price would result in small changes in quantity supplied and quantity demanded, and $(Q_S - Q_D)$ would be less than \$50 million as illustrated in Figure 9.5.a.ii.

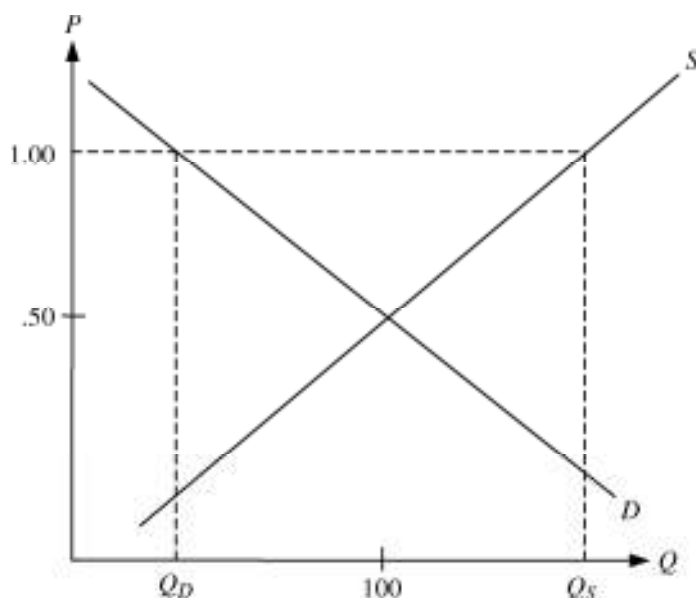


Figure 9.5.a.i

We can determine the combinations of supply and demand elasticities that yield either result. The elasticity of supply is $E_S = (\% \Delta Q_S) / (\% \Delta P)$, so the percentage change in quantity supplied is $\% \Delta Q_S = E_S (\% \Delta P)$. Since the price increase is 100% (from \$0.50 to \$1.00), $\% \Delta Q_S = 100 E_S$. Likewise, the percentage change in quantity demanded is $\% \Delta Q_D = 100 E_D$. The gap between Q_D and Q_S in percentage terms is $\% \Delta Q_S - \% \Delta Q_D = 100 E_S - 100 E_D = 100 (E_S - E_D)$. If this gap is exactly 50% of the current 100 million pounds of jelly beans, the gap will be 50 million pounds, and the cost of the price support program will be exactly \$50 million. So the program will cost \$50 million if $100(E_S - E_D) = 50$, or $(E_S - E_D) = 0.5$. If the difference between the elasticities is greater than one half, the program will

cost more than \$50 million, and if the difference is less than one half, the program will cost less than \$50 million. So the supply and demand can each be fairly inelastic (for example, 0.3 and -0.4) and still trigger a cost greater than \$50 million.

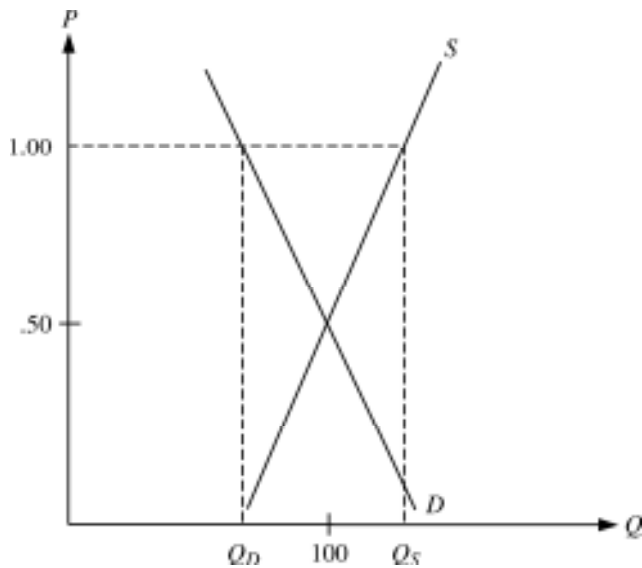


Figure 9.5.a.ii

- b. Could this program cost consumers (in terms of lost consumer surplus) *more* than \$50 million per year? Under what conditions? Could it cost consumers *less* than \$50 million per year? Under what conditions? Again, use a diagram to illustrate.

When the demand curve is perfectly inelastic, the loss in consumer surplus is \$50 million, equal to $(\$0.50)(100 \text{ million pounds})$. This represents the highest possible loss in consumer surplus, so the loss cannot be more than \$50 million per year. If the demand curve has any elasticity at all, the loss in consumer surplus will be less than \$50 million. In Figure 9.5.b, the loss in consumer surplus is area *A* plus area *B* if the demand curve is the completely inelastic *D* and only area *A* if the demand curve is *D'*.

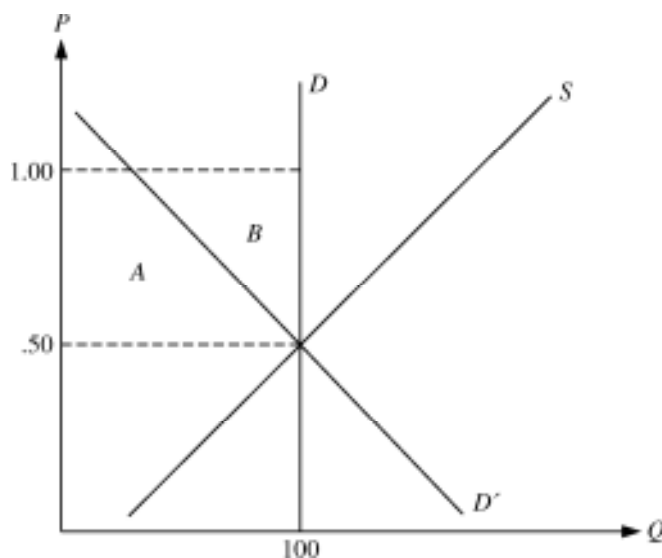


Figure 9.5.b

7. The United States currently imports all of its coffee. The annual demand for coffee by U.S. consumers is given by the demand curve $Q = 250 - 10P$, where Q is quantity (in millions of pounds) and P is the market price per pound of coffee. World producers can harvest and ship coffee to U.S. distributors at a constant marginal (= average) cost of \$8 per pound. U.S. distributors can in turn distribute coffee for a constant \$2 per pound. The U.S. coffee market is competitive. Congress is considering a tariff on coffee imports of \$2 per pound.

- a. If there is no tariff, how much do consumers pay for a pound of coffee? What is the quantity demanded?

If there is no tariff then consumers will pay \$10 per pound of coffee, which is found by adding the \$8 that it costs to import the coffee plus the \$2 that it costs to distribute the coffee in the United States. In a competitive market, price is equal to marginal cost. At a price of \$10, the quantity demanded is 150 million pounds.

- b. If the tariff is imposed, how much will consumers pay for a pound of coffee? What is the quantity demanded?

Now add \$2 per pound tariff to marginal cost, so price will be \$12 per pound, and quantity demanded is $Q = 250 - 10(12) = 130$ million pounds.

- c. Calculate the lost consumer surplus.

Lost consumer surplus is $(12 - 10)(130) + 0.5(12 - 10)(150 - 130) = \280 million.

- d. Calculate the tax revenue collected by the government.

The tax revenue is equal to the tariff of \$2 per pound times the 130 million pounds imported. Tax revenue is therefore \$260 million.

- e. Does the tariff result in a net gain or a net loss to society as a whole?

There is a net loss to society because the gain (\$260 million) is less than the loss (\$280 million).

9. Among the tax proposals regularly considered by Congress is an additional tax on distilled liquors. The tax would not apply to beer. The price elasticity of supply of liquor is 4.0, and the price elasticity of demand is -0.2 . The cross-elasticity of demand for beer with respect to the price of liquor is 0.1.

- a. If the new tax is imposed, who will bear the greater burden—liquor suppliers or liquor consumers? Why?

The fraction of the tax borne by consumers is given in Section 9.6 as $\frac{E_S}{E_S - E_D}$, where E_S is the own-price elasticity of supply and E_D is the own-price elasticity of demand. Substituting for E_S and E_D , the pass-through fraction is

$$\frac{4}{4 - (-0.2)} = \frac{4}{4.2} \approx 0.95.$$

Therefore, just over 95% of the tax is passed through to consumers because supply is highly elastic while demand is very inelastic. So liquor consumers will bear almost all the burden of the tax.

- b. Assuming that beer supply is infinitely elastic, how will the new tax affect the beer market?

With an increase in the price of liquor (from the large pass-through of the liquor tax), some consumers will substitute away from liquor to beer because the cross-elasticity is positive.

This will shift the demand curve for beer outward. With an infinitely elastic supply for beer (a horizontal supply curve), the equilibrium price of beer will not change, and the quantity of beer consumed will increase.

11. Example 9.6 (page 342) describes the effects of the sugar quota. In 2011, imports were limited to 6.9 billion pounds, which pushed the domestic price to 36 cents per pound. Suppose imports were expanded to 10 billion pounds.

a. What would be the new U.S. domestic price?

Example 9.6 gives equations for the total market demand for sugar in the U.S. and the supply of U.S. producers:

$$Q_D = 29.73 - 0.19P$$

$$Q_S = -7.95 + 0.66P.$$

The difference between the domestic quantities demanded and supplied, $Q_D - Q_S$, is the amount of imported sugar that is restricted by the quota. If the quota is increased to 10 billion pounds, then $Q_D - Q_S = 10$ and we can solve for P :

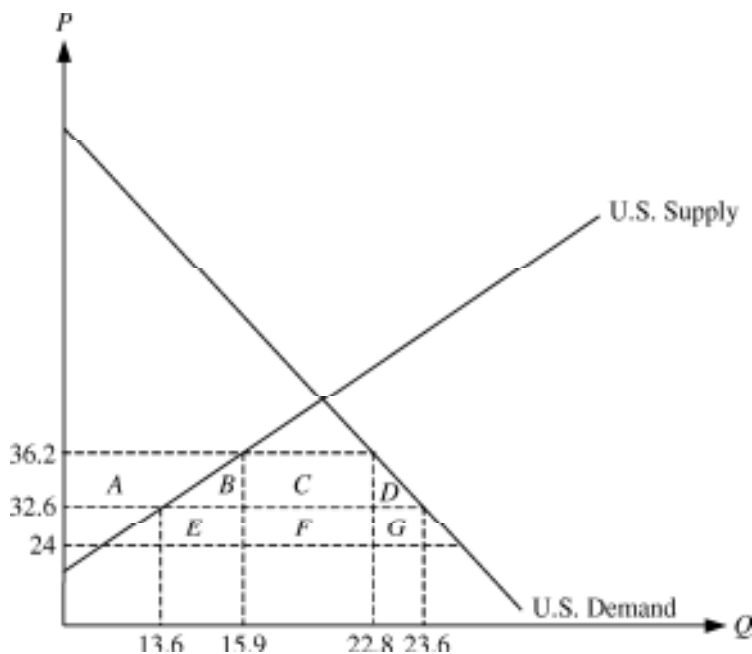
$$(29.73 - 0.19P) - (-7.95 + 0.66P) = 10$$

$$37.68 - 0.85P = 10$$

$$P = 32.6 \text{ cents per pound.}$$

At a price of 32.6 cents per pound, $Q_S = -7.95 + 0.66(32.6) = 13.6$ billion pounds, and $Q_D = Q_S + 10 = 13.6 + 10 = 23.6$ billion pounds.

b. How much would consumers gain and domestic producers lose?



The gain in consumer surplus is $A + B + C + D$. The loss to domestic producers is area A . The areas in billions of cents (i.e., tens of millions of dollars) are:

$$A = (13.6)(36.2 - 32.6) + (0.5)(15.9 - 13.6)(36.2 - 32.6) = 53.10$$

$$B = (0.5)(15.9 - 13.6)(36.2 - 32.6) = 4.14$$

$$C = (22.8 - 15.9)(36.2 - 32.6) = 24.84$$

$$D = (0.5)(23.6 - 22.8)(36.2 - 32.6) = 1.44$$

Thus, consumer surplus increases by 83.52, or \$835.2 million, while domestic producer surplus decreases by 53.1, or \$531 million.

c. What would be the effect on deadweight loss and foreign producers?

Domestic deadweight loss decreases by the difference between the increase in consumer surplus and the decrease in producer surplus, which is $\$835.2 - \$531.0 = \$304.2$ million.

When the quota was 6.9 billion pounds, the profit earned by foreign producers was the difference between the domestic price and the world price ($36.2 - 24$) times the 6.9 billion units sold, for a total of 84.18, or \$841.8 million. When the quota is increased to 10 billion pounds, domestic price falls to 32.6 cents per pound, and profit earned by foreigners is $(32.6 - 24)(10) = 86$, or \$860 million. Profit earned by foreigners therefore increases by \$18.2 million. On the diagram above, this is area $(E + F + G) - (C + F) = E + G - C$. The deadweight loss of the quota, including foreign producer surplus, decreases by area $B + D + E + G$. Area $E = 19.78$ and $G = 6.88$, so the decrease in deadweight loss $= 4.14 + 1.44 + 19.78 + 6.88 = 32.24$, or \$322.4 million.

13. Currently, the social security payroll tax in the United States is evenly divided between employers and employees. Employers must pay the government a tax of 6.2% of the wages they pay, and employees must pay 6.2% of the wages they receive. Suppose the tax were changed so that employers paid the full 12.4% and employees paid nothing. Would employees then be better off?

If the labor market is competitive (i.e., both employers and employees take the wage as given), then shifting all the tax onto employers will have no effect on the amount of labor employed or on employees' after tax wages. We know this because the incidence of a tax is the same regardless of who officially pays it. As long as the total tax doesn't change, the same amount of labor will be employed, and the wages paid by employers and received by employees (after tax) will not change. Hence, employees would be no better or worse off if employers paid the full amount of the social security tax.

15. In 2011, Americans smoked 16 billion packs of cigarettes. They paid an average retail price of \$5.00 per pack.

a. Given that the elasticity of supply is 0.5 and the elasticity of demand is -0.4, derive linear demand and supply curves for cigarettes.

Let the demand curve be of the general linear form $Q = a - bP$ and the supply curve be $Q = c + dP$, where a , b , c , and d are positive constants that we have to find from the information given above. To begin, recall the formula for the price elasticity of demand

$$E_P^D = \frac{P}{Q} \frac{\Delta Q}{\Delta P}.$$

We know the values of the elasticity, P , and Q , which means we can solve for the slope, which is $-b$ in the above formula for the demand curve.

$$\begin{aligned} -0.4 &= \left(\frac{5.00}{16} \right) (-b) \\ b &= 0.4 \left(\frac{16}{5.00} \right) = 1.28. \end{aligned}$$

To find the constant a , substitute for Q , P , and b in the demand curve formula: $16 = a - 1.28(5.00)$. Solving yields $a = 22.4$. The equation for demand is therefore $Q = 22.4 - 1.28P$. To find the supply curve, recall the formula for the elasticity of supply and follow the same method as above:

$$\begin{aligned} E_P^S &= \frac{P}{Q} \frac{\Delta Q}{\Delta P} \\ 0.5 &= \left(\frac{5.00}{16} \right) (d) \\ d &= 0.5 \left(\frac{16}{5.00} \right) = 1.6 \end{aligned}$$

To find the constant c , substitute for Q , P , and d in the supply formula, which yields $16 = c + 1.6(5.00)$. Therefore $c = 8$, and the equation for the supply curve is $Q = 8 + 1.6P$.

- b. Cigarettes are subject to a federal tax, which was about \$1.00 per pack in 2011. What does this tax do to the market-clearing price and quantity?**

The tax drives a wedge between supply and demand. At the new equilibrium, the price buyers pay, P_b , will be \$1.00 higher than the price sellers receive, P_s . Also, the quantity buyers demand at P_b must equal the quantity supplied at price P_s . These two conditions are:

$$P_b - P_s = 1.00 \quad \text{and} \quad 22.4 - 1.28P_b = 8 + 1.6P_s.$$

Solving these simultaneously, $P_s = \$4.56$ and $P_b = \$5.56$. The new quantity will be $Q = 22.4 - 1.28(5.56) = 15.3$ billion packs. So the price consumers pay will increase from \$5.00 to \$5.56 (a 56-cent increase) and consumption will fall from 16 to 15.3 billion packs per year (a drop of 700 million packs per year).

- c. How much of the federal tax will consumers pay? What part will producers pay?**

Consumers pay $\$5.56 - \$5.00 = \$0.56$ and producers pay the remaining $\$5.00 - \$4.56 = \$0.44$ per pack. We could also find these amounts using the pass-through formula. The fraction of the tax paid by consumers is $E_S/(E_S - E_D) = 0.5/[0.5 - (-0.4)] = 0.5/0.9 = 0.56$. Therefore, consumers will pay 56% of the \$1.00 tax, which is 56 cents, and suppliers will pay the remaining 44 cents.

Chapter 7: Uncertainty and Consumer Behavior

1. Consider a lottery with three possible outcomes:

- \$125 will be received with probability 0.2
- \$100 will be received with probability 0.3
- \$50 will be received with probability 0.5

a. What is the expected value of the lottery?

The expected value, EV , of the lottery is equal to the sum of the returns weighted by their probabilities:

$$EV = (0.2)(\$125) + (0.3)(\$100) + (0.5)(\$50) = \$80.$$

b. What is the variance of the outcomes?

The variance, σ^2 , is the sum of the squared deviations from the mean, \$80, weighted by their probabilities:

$$\sigma^2 = (0.2)(125 - 80)^2 + (0.3)(100 - 80)^2 + (0.5)(50 - 80)^2 = \$975.$$

c. What would a risk-neutral person pay to play the lottery?

A risk-neutral person would pay the expected value of the lottery: \$80.

3. Richard is deciding whether to buy a state lottery ticket. Each ticket costs \$1, and the probability of winning payoffs is given as follows:

Probability	Return
0.50	\$0.00
0.25	\$1.00
0.20	\$2.00
0.05	\$7.50

a. What is the expected value of Richard's payoff if he buys a lottery ticket? What is the variance?

The expected value of the lottery is equal to the sum of the returns weighted by their probabilities:

$$EV = (0.5)(0) + (0.25)(\$1.00) + (0.2)(\$2.00) + (0.05)(\$7.50) = \$1.025$$

The variance is the sum of the squared deviations from the mean, \$1.025, weighted by their probabilities:

$$\begin{aligned}\sigma^2 &= (0.5)(0 - 1.025)^2 + (0.25)(1 - 1.025)^2 + (0.2)(2 - 1.025)^2 + (0.05)(7.5 - 1.025)^2, \text{ or} \\ \sigma^2 &= 2.812.\end{aligned}$$

b. Richard's nickname is "No-Risk Rick" because he is an extremely risk-averse individual. Would he buy the ticket?

An extremely risk-averse individual would probably not buy the ticket. Even though the expected value is higher than the price of the ticket, $\$1.025 > \1.00 , the difference is not enough to compensate Rick for the risk. For example, if his wealth is \$10 and he buys a \$1.00 ticket, he would have \$9.00, \$10.00, \$11.00, and \$16.50, respectively, under the four possible outcomes. If his utility function is $U = W^{0.5}$, where W is his wealth, then his expected utility is:

$$EU = (0.5)(9^{0.5}) + (0.25)(10^{0.5}) + (0.2)(11^{0.5}) + (0.05)(16.5^{0.5}) = 3.157.$$

This is less than 3.162, which is his utility if he does not buy the ticket ($U(10) = 10^{0.5} = 3.162$). Therefore, he would not buy the ticket.

- c. **Richard has been given 1000 lottery tickets. Discuss how you would determine the smallest amount for which he would be willing to sell all 1000 tickets.**

With 1000 tickets, Richard's expected payoff is \$1025. He does not pay for the tickets, so he cannot lose money, but there is a wide range of possible payoffs he might receive ranging from \$0 (in the extremely unlikely event that all 1000 tickets pay nothing) to \$7500 (in the even more unlikely case that all 1000 tickets pay the top prize of \$7.50), and virtually everything in between. Given this variability and Richard's high degree of risk aversion, we know that Richard would be willing to sell all the tickets for less (and perhaps considerably less) than the expected payoff of \$1025. More precisely, he would sell the tickets for \$1025 minus his risk premium. To find his selling price, we would first have to calculate his expected utility for the lottery winnings. This would be like point *F* in Figure 5.4 in the text, except that in Richard's case there are thousands of possible payoffs, not just two as in the figure. Using his expected utility value, we then would find the certain amount that gives him the same level of utility. This is like the \$16,000 income at point *C* in Figure 5.4. That certain amount is the smallest amount for which he would be willing to sell all 1000 lottery tickets.

- d. **In the long run, given the price of the lottery tickets and the probability/return table, what do you think the state would do about the lottery?**

Given the price of the tickets, the sizes of the payoffs and the probabilities, the lottery is a money loser for the state. The state loses $\$1.025 - 1.00 = \0.025 (two and a half cents) on every ticket it sells. The state must raise the price of a ticket, reduce some of the payoffs, raise the probability of winning nothing, lower the probabilities of the positive payoffs, or some combination of the above.

5. **You are an insurance agent who must write a policy for a new client named Sam. His company, Society for Creative Alternatives to Mayonnaise (SCAM), is working on a low-fat, low-cholesterol mayonnaise substitute for the sandwich-condiment industry. The sandwich industry will pay top dollar to the first inventor to patent such a mayonnaise substitute. Sam's SCAM seems like a very risky proposition to you. You have calculated his possible returns table as follows:**

Probability	Return	Outcome
0.999	-\$ 1,000,000	(he fails)
0.001	\$ 1,000,000,000	(he succeeds and sells his formula)

- a. **What is the expected return of Sam's project? What is the variance?**

The expected return, ER , of Sam's investment is

$$ER = (0.999)(-1,000,000) + (0.001)(1,000,000,000) = \$1000.$$

The variance is

$$\sigma^2 = (0.999)(-1,000,000 - 1000)^2 + (0.001)(1,000,000,000 - 1000)^2, \text{ or}$$

$$\sigma^2 = 1,000,998,999,000,000.$$

b. What is the most that Sam is willing to pay for insurance? Assume Sam is risk neutral.

Suppose the insurance guarantees that Sam will receive the expected return of \$1000 with certainty regardless of the outcome of his SCAM project. Because Sam is risk neutral and because his expected return is the same as the guaranteed return with insurance, the insurance has no value to Sam. He is just as happy with the uncertain SCAM profits as with the certain outcome guaranteed by the insurance policy. So Sam will not pay anything for the insurance.

c. Suppose you found out that the Japanese are on the verge of introducing their own mayonnaise substitute next month. Sam does not know this and has just turned down your final offer of \$1000 for the insurance. Assume that Sam tells you SCAM is only six months away from perfecting its mayonnaise substitute *and* that you know what you know about the Japanese. Would you raise or lower your policy premium on any subsequent proposal to Sam? Based on his information, would Sam accept?

The entry of the Japanese lowers Sam's probability of a high payoff. For example, assume that the probability of the billion-dollar payoff cut in half. Then the expected outcome is:

$$ER = (0.9995)(-\$1,000,000) + (0.0005)(\$1,000,000,000) = -\$499,500.$$

Therefore you should raise the policy premium substantially. But Sam, not knowing about the Japanese entry, will continue to refuse your offers to insure his losses.

7. Suppose that two investments have the same three payoffs, but the probability associated with each payoff differs, as illustrated in the table below:

Payoff	Probability (Investment A)	Probability (Investment B)
\$300	0.10	0.30
\$250	0.80	0.40
\$200	0.10	0.30

a. Find the expected return and standard deviation of each investment.

The expected value of the return on investment A is

$$EV = (0.1)(300) + (0.8)(250) + (0.1)(200) = \$250.$$

The variance on investment A is

$$\sigma^2 = (0.1)(300 - 250)^2 + (0.8)(250 - 250)^2 + (0.1)(200 - 250)^2 = 500,$$

and the standard deviation on investment A is $\sigma = \sqrt{500} = \$22.36$.

The expected value of the return on investment B is

$$EV = (0.3)(300) + (0.4)(250) + (0.3)(200) = \$250.$$

The variance on investment B is

$$\sigma^2 = (0.3)(300 - 250)^2 + (0.4)(250 - 250)^2 + (0.3)(200 - 250)^2 = 1500,$$

and the standard deviation on investment B is $\sigma = \sqrt{1500} = \$38.73$.

- b. Jill has the utility function $U = 5I$, where I denotes the payoff. Which investment will she choose?**

Jill's expected utility from investment A is

$$EU = (0.1)(5 \times 300) + (0.8)(5 \times 250) + (0.1)(5 \times 200) = 1250.$$

Jill's expected utility from investment B is

$$EU = (0.3)(5 \times 300) + (0.4)(5 \times 250) + (0.3)(5 \times 200) = 1250.$$

Since both investments give Jill the same expected utility she will be indifferent between the two. Note that Jill is risk neutral, so she cares only about expected values. Since investments A and B have the same expected values, she is indifferent between them.

- c. Ken has the utility function $U = 5\sqrt{I}$. Which investment will he choose?**

Ken's expected utility from investment A is

$$EU = (0.1)(5)\sqrt{300} + (0.8)(5)\sqrt{250} + (0.1)(5)\sqrt{200} = 78.98.$$

Ken's expected utility from investment B is

$$EU = (0.3)(5)\sqrt{300} + (0.4)(5)\sqrt{250} + (0.3)(5)\sqrt{200} = 78.82.$$

Ken will choose investment A because it has a slightly higher expected utility. Notice that Ken is risk averse, and since the two investments have the same expected return, he prefers the investment with less variability.

- d. Laura has the utility function $U = 5I^2$. Which investment will she choose?**

Laura's expected utility from investment A is

$$EU = (0.1)(5 \times 300^2) + (0.8)(5 \times 250^2) + (0.1)(5 \times 200^2) = 315,000.$$

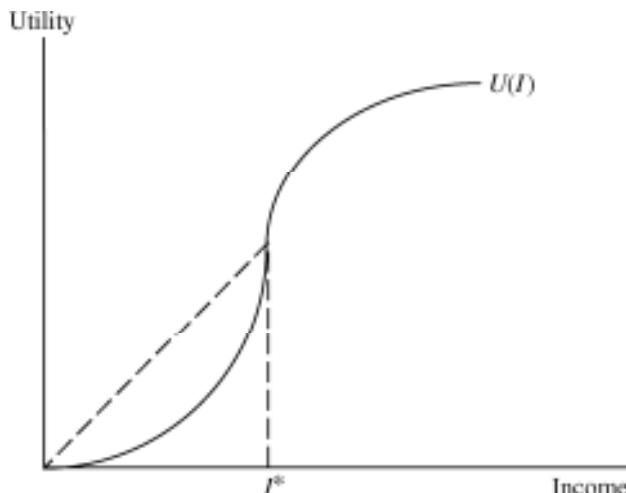
Laura's expected utility from investment B is

$$EU = (0.3)(5 \times 300^2) + (0.4)(5 \times 250^2) + (0.3)(5 \times 200^2) = 320,000.$$

Laura will choose investment B since it has a higher expected utility. Notice that Laura is a risk lover, and since the two investments have the same expected return, she prefers the investment with greater variability.

- 9. Draw a utility function over income $u(I)$ that describes a man who is a risk lover when his income is low but risk averse when his income is high. Can you explain why such a utility function might reasonably describe a person's preferences?**

The utility function will be S-shaped as illustrated below. Preferences might be like this for an individual who needs a certain level of income, I^* , in order to stay alive. An increase in income above I^* will have diminishing marginal utility. Below I^* , the individual will be a risk lover and will take unfavorable gambles in an effort to make large gains in income. Above I^* , the individual will purchase insurance against losses and below I^* will gamble.



11. A moderately risk-averse investor has 50% of her portfolio invested in stocks and 50% in risk-free Treasury bills. Show how each of the following events will affect the investor's budget line and the proportion of stocks in her portfolio:

- a. The standard deviation of the return on the stock market increases, but the expected return on the stock market remains the same.

From Section 5.4, the equation for the budget line is

$$R_p = \left[\frac{R_m - R_f}{\sigma_m} \right] \sigma_p + R_f,$$

where R_p is the expected return on the portfolio, R_m is the expected return from investing in the stock market, R_f is the risk-free return on Treasury bills, σ_m is the standard deviation of the return from investing in the stock market, and σ_p is the standard deviation of the return on the portfolio. The budget line is linear and shows the positive relationship between the return on the portfolio, R_p , and the standard deviation of the return on the portfolio, σ_p , as shown in Figure 5.6.

In this case σ_m , the standard deviation of the return on the stock market, increases. The slope of the budget line therefore decreases, and the budget line becomes flatter. The budget line's intercept stays the same because R_f does not change. Thus, at any given level of portfolio return, the portfolio now has a higher standard deviation. Since stocks have become riskier without a compensating increase in expected return, the proportion of stocks in the investor's portfolio will fall.

- b. The expected return on the stock market increases, but the standard deviation of the stock market remains the same.

In this case, R_m , the expected return on the stock market, increases, so the slope of the budget line becomes steeper. At any given level of portfolio standard deviation, σ_p , there is now a higher expected return, R_p . Stocks have become relatively more attractive because investors now get greater expected returns with no increase in risk, and the proportion of stocks in the investor's portfolio will rise as a consequence.

- c. The return on risk-free Treasury bills increases.

In this case there is an increase in R_f , which affects both the intercept and slope of the budget line. The budget line shifts up and becomes flatter as a result. The proportion of stocks in the

portfolio could go either way. On one hand, Treasury bills now have a higher return and so are more attractive. On the other hand, the investor can now earn a higher return from each Treasury bill and so could hold fewer Treasury bills and still maintain the same level of risk-free return. In this second case, the investor may be willing to place more of her money in the stock market. It will depend on the particular preferences of the investor as well as the magnitude of the returns to the two asset classes. An analogy would be to consider what happens to savings when the interest rate increases. On the one hand, savings tend to increase because the return is higher, but on the other hand, spending may increase and savings decrease because a person can save less each period and still wind up with the same accumulation of savings at some future date.

Chapter 13 The United States in a Global Economy

■ Answers to Odd-End-of-Chapter Questions

1. How can globalization and international economic integration be measured?

Answer: The chapter offers three ways to measure globalization and economic integration: (1) trade flows; (2) factor movements; and (3) convergence of prices (goods, factors, and assets).

3. What does the trade-to-GDP ratio measure? Does a low value indicate that a country is closed to trade with the outside world?

Answer: The trade-to-GDP ratio is a measure of the relative importance of trade to a national economy. It is measured by the ratio of exports plus imports to GDP.

A relatively small ratio does not necessarily mean that an economy is intentionally closed to the outside world. Large countries like the United States have large domestic markets that enable firms to specialize and produce in volume in order to attain an optimal scale. Specialization and high volume in manufacturing is often associated with increased productivity, so firms in large markets can achieve the highest possible level of productivity without having to sell to foreign markets. Firms located in smaller countries have to trade their output across international boundaries if they want to have the same technology and the same level of productivity. Consequently, large countries tend to have lower trade-to-GDP ratios regardless of their trade policies.

5. Trade and capital flows were described and measured in relative rather than absolute terms. Explain the difference. Which term seems more valid—*relative* or *absolute*? Why?

Answer: Absolute values are the dollar amounts of trade and capital flows. Relative values are the ratio of dollar values to GDP. Relative values are a better indicator of the importance of trade and capital flows since they are proportional to the size of national economies. Large economies like the United States may have large export and import values, but the importance of trade to the national economy is not nearly as great as it is for other economies. The United States is a large exporter and importer, but the national economy is so large that trade is much less important for the United States than it is for many smaller countries such as Canada, Belgium, or The Netherlands.

7. What are the new issues in international trade and investment? In what sense do they expose national economies to outside influences?

Answer: The new issues involve policy differences between nations that until recently were considered the exclusive responsibility of local or national governments. Examples include labor standards, environmental standards, competition or antitrust policies, and industrial support policies.

Negotiations between nations potentially give foreign interests a voice in setting domestic policy. The scope and the depth of the negotiations determine how great a voice foreigners will have. It is often the case, however, that negotiations either occur or are proposed because some aspect of domestic policy is perceived by foreigners as a barrier to trade, and they seek to alter the domestic policy that created it.

Chapter 14 Comparative Advantage and the Gains from Trade

■ Answers to Odd-End-of-Chapter Questions

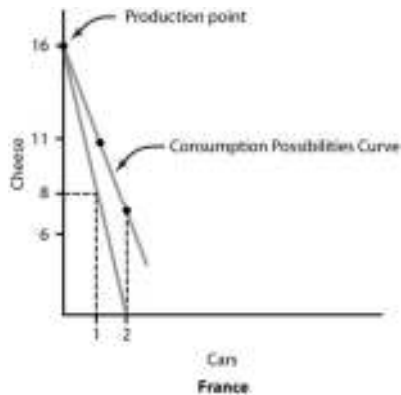
1. Use the information in the table on labor productivities in France and Germany to answer questions a through f.

	Output per Hour Worked	
	France	Germany
Cheese	2 kilograms	1 kilogram
Cars	0.25	0.5

- Which country has an absolute advantage in cheese? In cars?
- What is the relative price of cheese in France if it does not trade? In Germany?
- What is opportunity cost of cheese in France? In Germany?
- Which country has a comparative advantage in cheese? In cars? Explain your answer.
- What are the upper and lower bounds for the trade price of cheese?
- Draw a hypothetical PPC for France and label its slope. Suppose that France follows its comparative advantage in deciding where to produce on its PPC. Label its production point. If the trade price of cars is 5 kilograms of cheese per car, draw a trade line (CPC) showing how France can gain from trade.

Answers:

- France has the absolute advantage in cheese and Germany in cars. This follows because France's productivity is higher in cheese and Germany's is higher in cars.
- The autarkic relative price of cheese in France is one-eighth car per kilogram; in Germany it is one-half car.
- Opportunity costs are equal to relative prices.
- France has a comparative advantage in cheese because its opportunity cost is lower (one-eighth car versus one-half car in Germany). By the same reasoning, Germany has a comparative advantage in cars.
- The trade price of cheese will settle between one-eighth and one-half car per kilo.
-



3. Explain how a nation can gain from trade even though not everyone is made better off. Is this a contradiction?

Answer: This is not a contradiction. The gains from trade imply that the winners could compensate the losers completely and still have gains left over. Some people may lose jobs (in the example in the text, as the United States moved toward specialization in steel, American workers were shifted out of the bread industry and into the steel industry), but others benefit from the higher demand for their product (workers in the steel industry). As long as the winner's gains are greater than the loser's losses, we can conclude that the nation wins.

5. Many people believe that the goal of international trade should be to create jobs. Consequently, when they see workers laid off due to a firm's inability to compete against cheaper and better imports, they assume that trade must be bad for the economy. Is this assumption correct? Why, or why not?

Answer: The goal of trade is to improve a nation's allocation of its resources so they are directed to their most valuable use. Trade is not about creating jobs, but is about raising the standard of living through a more efficient allocation of resources. Trade may cause workers to become laid off if they are in inefficient industries that do not produce according to the national comparative advantage. While this may be hard on the people who lose their jobs, it also frees up labor and capital so it can be directed to better uses.

Chapter 15 Exchange Rates and Exchange Rate Systems

■ Answers to Odd-End-of-Chapter Questions

1. Draw a graph of the supply of and demand for the Canadian dollar by the U.S. market. Diagram the effect of each of the following on exchange rates, state in words whether the effect is long, medium, or short run, and explain your reasoning.
 - a. More rapid growth in Canada than in the United States.
 - b. A rise in U.S. interest rates.
 - c. Goods are more expensive in Canada than in the United States.
 - d. A recession in the United States.
 - e. Expectations of a future depreciation in the Canadian dollar.

Answers:

- a. The Canadian supply of currency to the U.S. market increases in response to the rise in Canada's demand for American exports. The supply curve shifts right; the U.S. dollar appreciates; the Canadian dollar depreciates. This effect is medium run because effects of economic expansions and contractions—the business cycle—on exchange rates run for a few years (usually less than a decade). As Canada experiences more rapid economic growth than the United States, disposable income in Canada rises, causing consumption to rise. Consumer confidence gradually rises as jobs become secured and plentiful. Canadian expenditures on imports rise.
- b. The supply of Canadian dollars to the U.S. market increases in response to the higher interest rates; the supply curve shifts right; the U.S. dollar appreciates; the Canadian dollar depreciates. This is a short-run effect because the factors that cause interest rates to change are themselves short-run processes. Good examples are changes in government (fiscal and/or monetary) policies, expectations, and financial capital flows in and out of a country.
- c. The U.S. demand for Canadian dollars decreases in response to higher prices for Canadian goods. The demand curve shifts left causing the exchange rate to fall. The U.S. dollar appreciates and the Canadian dollar depreciates. This is a long-run effect in part because the prices of goods move gradually. The higher prices of goods in Canada could also be caused by government policies such as tariffs and quotas. In general, changing government policies takes time. Goods arbitrage will eventually equalize the prices of goods between the two countries, but achieving purchasing power parity is a long-run process because of the following factors: (1) shipping, insurance, and other transportation may be prohibitively expensive; (2) trade barriers such as tariffs, quotas, import license, and inspection fees may be too high; and (3) a substantial number of goods may not be traded. All of these play a significant factor for purchasing power parity to exert influence only in the long run.
- d. The U.S. demand for Canadian dollars decreases in response to the drop in demand for imports; the demand curve shifts left; the U.S. dollar appreciates; the Canadian dollar depreciates. As explained in (a) above, the effect of recessions and expansions on exchange rates can be considered medium term.
- e. The demand for Canadian dollars decreases in response to its expected loss in value; the demand curve shifts left; the U.S. dollar appreciates; the Canadian dollar depreciates. Like the effects of interest rates, the effects of expectations on exchange rates are short run. As everyone knows, expectations could change swiftly and could reverse course almost instantaneously. Some

expectations could be unexpected and flimsy but they could have catastrophic effects on the exchange rates and financial sectors of the country.

See Table 10.3 for a summary of the short-, medium- and long-run factors that determine exchange rates.

3. Suppose the dollar-yen exchange rate is 0.01 dollar per yen. Since the base year, inflation has been 2 percent in Japan and 10 percent in the United States. What is the real exchange rate? In real terms, has the dollar appreciated or depreciated against the yen?

Answer: The real exchange rate is $R_r = 0.01(102/110) = 0.0093$ dollar per yen. The dollar has appreciated.

5. If U.S. visitors to Mexico can buy more goods in Mexico than they can in the United States when they convert their dollars to pesos, is the dollar undervalued or overvalued? Explain.

Answer: The dollar is overvalued and the peso is undervalued. The dollar buys “too many” pesos when it is converted. Hence it buys more in Mexico after its conversion to pesos. Conversely, a traveler to the United States would find that the pesos he exchanged for dollars buys him fewer goods than the same pesos spent in Mexico.

7. In the debate on fixed versus floating exchange rates, the strongest argument for a floating rate is that it frees macroeconomic policy from taking care of the exchange rate. This is also the weakest argument. Explain.

Answer: Fixed exchange rate systems require the monetary authority to closely monitor the exchange rate. In effect, the domestic money supply is a captive of the need to maintain sufficient reserves to be able to supply any excess demand for foreign exchange. The potential conflict in this arrangement is that the needs of the exchange rate system can be in conflict with the needs of the domestic economy. This is the scenario that the United States and the United Kingdom faced during the Great Depression of the 1930s. Interest rates were raised in order to reduce the demand for foreign exchange; the rise in interest rates deepened the recession and caused unemployment to rise further.

Nevertheless, the freeing of monetary policy from the task of maintaining an exchange rate has its own problems. Some economists believe that the lack of external discipline on monetary policy leads to an overreliance on inflationary policies to satisfy domestic economic needs. Argentina, for example, was unable to cure its constant tendency toward hyperinflation until it abandoned the ability to freely change the money supply.

9. Suppose that U.S. interest rates are 4 percent more than rates in the EU.
 - a. Would you expect the dollar to appreciate or depreciate against the euro, and by how much?
 - b. If, contrary to your expectations, the forward and spot rates are the same, which direction would you expect financial capital to flow? Why?

Answers:

- a. Capital would flow into the United States increasing the supply of foreign exchange. Due to higher interest rates, investment at home is more attractive than in Europe. This also reduces the demand for foreign exchange. As a result, we expect the dollar to appreciate by 4 percent. This interest rate arbitrage activity continues until equilibrium is restored (that is, interest rate parity

is
reestablished).

- b. This suggests that the right-hand side of the interest parity equation is equal to zero, while the left-hand side of the same equation is greater than zero. If $F = R$ as the statement suggests, then currency markets are signaling that no changes are expected in the exchange rate. Capital would flow to the United States, decreasing the demand for foreign currency and increasing the supply of foreign currency. Both of these decrease R .